Project Number: **770299**

NewsEye:

A Digital Investigator for Historical Newspapers

Research and Innovation Action
Call H2020-SC-CULT-COOP-2016-2017

D4.7: Intelligible representation of statistical analysis (b) (final)

Due date of deliverable: M45 (31 January 2021)

Actual submission date: 31 January 2021

Start date of project: 1 May 2018 **Duration:** 45 months

Partner organization name in charge of deliverable: UH-CS

Project co-funded by the European Commission within Horizon 2020		
Dissemination Level		
PU	Public	PU
PP	Restricted to other programme participants (including the Commission Services)	-
RE	Restricted to a group specified by the Consortium (including the Commission Services)	-
CO	Confidential, only for members of the Consortium (including the Commission Services)	-

Revision History

Document administrative information		
Project acronym:	NewsEye	
Project number:	770299	
Deliverable number:	D4.7	
Deliverable full title:	Intelligible representation of statistical analysis (b) (final)	
Deliverable short title:	Intelligible representation (final)	
Document identifier:	NewsEye-T43-D47-IntelligibleRepresentation-b-Submitted-v6.0	
Lead partner short name:	UH-CS	
Report version:	V6.0	
Report preparation date:	31.01.2021	
Dissemination level:	PU	
Nature:	Report	
Lead author:	Elaine Zosa (UH-CS)	
Co-authors:	Michele Boggia (UH-CS), Lidia Pivovarova (UH-CS), Mark Granroth-Wilding (UH-CS), Hai Nguyen (ULR)	
Internal reviewers:	Eva Pfanzelter (UIBK-ICH), Martin Gasteiner (UNIVIE)	
Status:		Draft
		Final
	x	Submitted

The NewsEye Consortium partner responsible for this deliverable has addressed all comments received, making changes as necessary. Changes to this document are detailed in the change log table below.

Change Log

Date	Version	Editor	Summary of changes made
25/03/2021	0.1	Elaine Zosa (UH-CS) & Michele Boggia (UH-CS) & Hai Nguyen (ULR)	First outline draft
10/04/2021	1.0	Elaine Zosa (UH-CS) & Michele Boggia (UH-CS)	Prepared for internal review
22/04/2021	2.0	Mark Granroth-Wilding (UH-CS)	Reviews taken into account and submission for quality management (QM)
29/04/2021	2.1	Mark Granroth-Wilding (UH-CS)	Further updates following QM
30/04/2021	3.0	Antoine Doucet (ULR)	Minor adjustments and submission
03/12/2021	4.0	Elaine Zosa (UH-CS)	Draft update including works of the project extension period
04/01/2022	5.0	Elaine Zosa (UH-CS)	Final update following internal reviews
31/01/2022	6.0	Antoine Doucet (ULR)	Minor adjustments and submission

Executive summary

The NewsEye project addresses challenges relating to the exploration of historical news corpora. It makes contributions in text recognition, text analysis, natural language processing (NLP) and generation (NLG); in digital newspaper research; in digital humanities; and in history, in terms of analysing historical assets with new methods.

WP4, entitled 'Dynamic text analysis', aims to develop and implement methods for *contextualised* and *contrastive content analysis*, carried out *dynamically*, both for use directly by the demonstrator and by the autonomous investigator. Effective methods to display the analysis results in an intelligible form to the user are essential to the practical use of complex analysis methods. In this task, *T4.3: Intelligible Representations of Statistical Analysis*, we are developing methods and tools for presenting results for analysis methods developed in T4.1 (Analysis of content in a given context), primarily *topic models* (TMs).

The methods for automatic topic labelling aim to generate concise labels (or descriptors) of topics to provide users with an idea of what a topic is about and how the topics in the model differ from each other. In the case of dynamic topics, they should also tell the user how the topic changed over time. We also investigate methods for generating more detailed topic descriptions using extractive multi-document summarisation methods. For topic model visualisations, we want to provide users with a visual overview of a topic and a way to explore the different topics in the model through interactive ways. For dynamic topic models, we also want to show how a topic's prominence changes over time and provide a picture of a topic's evolution.

We report on the work we have done to generate topic labels and descriptions and topic model visualisations, using a range of existing methods and developments of them, with experiments to compare their effectiveness. We document their integration into the NewsEye pipeline. These tools are available both for the end users and for the automated *Personal Research Assistant* through a REST API that provides access to the tools developed in the tasks of WP4.

Contents

1. Introduction

In this section, we set the work of WP4 in the broader context of NewsEye, describe the goals of Task T4.3 and summarise the work carried out for this task.

1.1. Context within NewsEye

The NewsEye project addresses a number of challenges relating to exploration of historical news corpora. These involve contribution in several directions:

- in text recognition, text analysis, natural language processing, computational creativity and natural language generation, particularly with regard to historical newspapers;
- in digital newspaper research, addressing a number of editorial issues like OCR and article separation;
- in digital humanities, dealing with huge amounts of text material, availability of useful tools and possibilities of searching and browsing; and
- in history, in terms of analysing historical assets with new methods across different language corpora.

Central to the project are the *Demonstrator*, a means for a user to explore large collections, and the *Personal Research Assistant* (PRA), a tool to perform autonomous exploratory search of collections to help a user identify content of interest. The PRA consists of the *Investigator*, carrying out autonomous analysis, the *Reporter*, delivering reports on the results to the user, and the *Explainer*, explaining how the results were arrived at and why they may be of interest. The interactions between these components are described by Figure 1.

At the heart of both the Demonstrator and the Investigator component of the PRA lies a collection of tools for analysing historical newspaper data, made available in textual form by WP2 (Text Recognition and Article Separation) and enhanced with semantic annotations by WP3 (Semantic Text Enrichment).

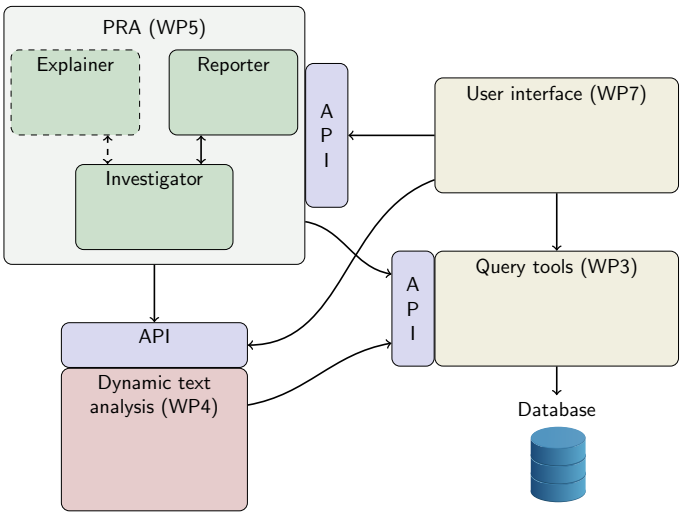


Figure 1: High-level architecture of component systems of NewsEye, showing how WP4 will interact with other WPs to acquire data and provide analysis.

WP4 provides a set of tools for broad-scale analysis of the collection and analysis of smaller groups of articles in the context of the whole collection. These tools will be used both by the user directly (through the Demonstrator) and by the autonomous Investigator.

1.2. Task T4.3

The main objective of WP4 is to develop and implement methods for contextualised and contrastive content analysis, carried out dynamically. In this task, *T4.3: Intelligible Representations of Statistical Analysis*, we are developing the methods and tools for performing this analysis, primarily using *topic models* (TMs).

In this deliverable, we report on the task of providing intelligible representations of statistical analysis of documents and corpora provided by the topic models developed in *Task T4.1*. The goals of these methods are:

- provide intelligible representations of topic analysis of a corpus or sub-corpus
- provide intelligible representations of the topic distributions of individual documents or set of documents
- provide topic labels that are descriptive of each topic and distinguish it from the other topics found by the topic model
- make the outputs of these methods available to users through the Personal Research Assistant using an API.

A variety of methods exist in the literature for representing topics and topic model analysis in different ways. We compare these methods experimentally to address the question of what the best methods are to present such information in the context of the NewsEye demonstrator.

We begin by briefly introducing the TMs that we focus on in this deliverable in Section 2. (A more detailed introduction to TMs, including other types of models used within NewsEye can be found in public deliverable D4.5) In Section 3 we describe some datasets that we use for training models in order to explore representation methods.

We use two approaches in the task of providing intelligible representations of the results of the analysis done in *Task T4.1*, specifically, the results of topic model training: the first is generating *textual topic descriptions* and the second is generating *visualisations* of topics and TMs. We divide the first approach into two sub-tasks: *topic labelling*, reported in Section 4 and *topic summarisation*, reported in Section 5.

For the visualisation approach, in Section 6, we report on techniques to represent a topic visually, such as word clouds and interactive plots. We also discuss some novel methods for presenting the temporal aspect of dynamic topic models, such as topic evolution and change in topic prominence. We use existing tools and libraries to generate these visualisations. Word clouds and LDAVis are already familiar methods for visualising static topic models while heatmaps and bar charts are not as commonly used when it comes to dynamic topic modelling.

All of the tools provided by WP4 are available to both the Demonstrator and the Investigator via a set of APIs. Section 7 details the API calls that other WPs can already use to access some of the methods described here.

In Section 8 we describe our collaboration with digital humanities researchers on how we used the visualisation methods in a paper investigating discourse dynamics in Finnish newspapers.

Section 9 describes the visualisation of the evolution of stances towards particular named entities (NEs).

In Section 10, we conclude with a summary of the work we have done for this task.

2. Introduction to topic models

In this section we provide a brief introduction on the topic models (TMs) used in this task. Since this task is about representation of TMs and their analyses, we will skip over the technical details of how these models infer topics and focus on what the models provide.

Standard TMs such as Latent Dirichlet Allocation (LDA) learn topics from a set of documents, where a *topic* is a probability distribution over the vocabulary. The words that have the highest probabilities in a topic's probability distribution are typically considered to be the most important words for that topic and therefore to tell us something about what the topic is about.

Static TMs such as LDA learn *static topics*, meaning that each topic has a single distribution over the vocabulary. In the case of documents with timestamps covering some time interval, such as news articles, we also want to capture dynamic co-occurrence patterns that evolve through time. Dynamic topic models [DTM, Blei and Lafferty, 2006] capture themes or topics discussed in a set of time-stamped documents and how the words related to these topics change in prominence with time. In DTM, the dataset is divided into *time slices* and the model infers topic distributions that evolve in each consecutive time slice.

LDA and DTM are monolingual TMs, meaning that they are applicable only to corpora composed of documents in a single language. We are also interested in corpora where documents can be in a variety of languages. Multilingual topic models are developed to capture cross-lingual topics from multilingual datasets. This means that topics learned from these models are aligned across languages. Polylingual topic model [PLTM, Mimno et al., 2009] is an extension of LDA that infers topics from an aligned multilingual corpus composed of document tuples. Tuples contain documents in one or more languages that are thematically aligned, for example describing the same event or discussing the same issue.

Aside from topic distributions over the vocabulary, TMs also estimate a probability distribution for each document over the topics. This tells us how much of each document is about a topic. Together, these two sets of distributions, which we will refer to as the *topic-term distributions* and *document-topic distributions*, will be used to analyse corpora and individual documents and present the output of that analysis to the user.

3. Datasets and trained models

In order to explore representation methods for TMs, we first require some trained models. Since we are investigating both monolingual models (LDA) and multilingual models (PLTM), we need multilingual datasets on which to train. Since we are also interested in *dynamic* TMs (DTM), we use datasets whose documents are timestamped and cover some extended period. Here we describe a number of suitable datasets and the models we trained on them.

3.1. NLF Finnish Dataset

The National Library of Finland (NLF) has made available digitized newspapers from Finland from 1790-1917[1]. Since this dataset continuously covers a long time period we use this for our experiments on dynamic topics.

We trained LDA and DTM models with 50 topics on the Finnish portion of this data for 64 years (1854-1917).

3.2. NewsEye French Dataset

We used a portion of the French newspapers in the NewsEye collection to develop some our topic labelling and topic description methods. Specifically we used articles from the year 1915 because the data is relatively clean and smaller than the Finnish and Austrian datasets.

We trained an LDA model for 30 topics on this dataset.

3.3. Reuters Dataset

The Reuters dataset is a collection of Reuters news articles published from 1996-1997. Each article includes the article body, headline, publication date and pre-assigned news categories the article belongs to.

We trained an LDA model for 50 topics with the Reuters dataset.

3.4. DE-News Dataset

The *DE-News* corpus is a German-English (*de-en*) parallel dataset of news articles from August 1996 to January 2000[2]. The articles are originally in German and have been manually translated into English. The fact that the articles are aligned across languages allows us to train PLTM on a set with reliable document alignments.

We trained a PLTM with 10 topics on this dataset. This means that we have 10 separate topic distributions, aligned across the two languages, giving us 20 topic distributions in all.

4. Topic labelling

In topic modelling, a *topic* is a probability distribution over a vocabulary. The widely used method of representing a topic for inspection is to present the words in the vocabulary that have a high probability in the topic-term distribution. This method produces mostly satisfactory results but is also presents some problems which we discuss below. In this section we present some of the alternative methods used to represent topics.

[1] https://digi.kansalliskirjasto.fi/opendata
[2] http://homepages.inf.ed.ac.uk/pkoehn/publications/de-news/

	Reuters LDA Topics
1	pakistan northern ireland irish lebanon
2	bonds municipal million inc county
3	canada canadian belgian beef cattle
4	bond million issue notes year
5	wheat prices per tonne corn
6	drug tobacco health medical lawsuit
7	japan yen japanese tokyo yuan
8	zealand test day first bomb
9	market points data percent bond
10	german french dollar france marks

Table 1: Most probable words for some topics from the Reuters LDA model

	French historical LDA Topics
1	rue saint gare mme train
2	loi police projet payer intérêts
3	journal septembre article suivant titre
4	entier pourtant civils mots combien
5	nombreux bois mètres maisons dernières
6	bons divers assurer désormais chevaux
7	elles trop chaque parce doute
8	théâtre palais père revue comédie
9	mer valdès marins côte appareil
10	guerre deux dont leurs entre

Table 2: Most probable words for some topics from the historical French LDA model

4.1. Word probability

The most commonly used method to inspect topics in LDA and related TMs is to display for each topic the most probable words in the topic-term distribution. Often, the top N words (where $N \simeq 10$) are displayed together with their probabilities $p(w|t)$.

A key problem with this method is that some words are simply common in all topics, since they appear frequently throughout the corpus. Whilst these words may have a high probability given the topic, they are not necessarily *characteristic* of the topic, and might not give a good reflection of the topic's content [Sievert and Shirley, 2014]. Tables 1 and 2 show the most topics probable words of topics from the LDA models trained on the historical French dataset and the modern Reuters dataset. Topic 8 in the Reuters topics is an example of a topic where several top words do not reflect the topic content. This topic contains the words 'first' and 'day', words that do not contribute much to the interpretability of the topic.

4.2. Lift

Sievert and Shirley [2014] use a metric *lift* defined by Taddy [2012] to rank words by their significance for a topic. The lift of a word w for a topic k is defined as:

Most probable	**Lift**	**Relevance**
1: pakistan northern ireland irish lebanon	ira pakistani lebanese beirut fein	pakistan irish ireland lebanon ira
2: bonds municipal million inc county	incorporated insured paineweber dtc gos	municpial bonds county approx school
3: canada canadian belgian beef cattle	beef philippine manila pork hog	canada canadian belgian beef philippine
4: bond million issue notes year	borrower iss denoms und hryvnia	notes bond mln maturity bln
5: wheat prices per tonne corn	soybean cbot soybeans soymeals bushels	wheat corn tonne soybean cbot
6: drug tobacco health medical lawsuit	tobacco lawsuit miami smoking cigarette	tobacco drug health lawsuit miami
7: japan yen japanese tokyo yuan	yuan shanghai shenzhen ramos rao	japan yen japanese tokyo yuan
8: zealand test day first bomb	cricket blast overs wickets	zealand test cricket bomb algeria
9: market points data percent bond	dow greenspan ftse cpi liffe	data dow fed bond points
10: german french dollar france marks	francs chirac kohl bonn waigel	german french marks francs france

Table 3: Top words from some topics from the Reuters LDA model using different measures.

$$p(w|k)/p(w)$$

where $p(w)$ is the probability of the word occurring independently of the topic, which can be estimated from its frequency across the whole training corpus.

This has the effect of giving a higher weight to words that have a higher probability for this topic than would be expected from their occurrence in other contexts and down-weights words that have a high probability in the entire corpus. In general, the effect of this would be to highlight the more *characteristic* words in a specific topic. One benefit of this is that it can effectively exclude very common words, such as stop words, which can be seen in some topics in Tables 1 and 2.

Tables 3 and 4 show top topic words measured by lift. For comparison, we also show most probable topic words. We see significant differences in the top words of some Reuters topics such as Topic 3 where lift upweights words related to the Philippines whereas topic probability emphasises words related to Canada. In Topic 10, topic probability focuses on currencies while lift gives higher weights to names of government officials. In the historical French topics, however, we do not see much difference between the top words using different measures. This is because in this particular dataset, the words with the highest probability for each topic also happen to be high-frequency words in the dataset.

Most probable	Lift	Relevance
1 : rue saint gare mme train	rue gare mme train ami	rue saint gare mme train
2 : loi police projet payer intérêts	loi police projet payer intérêts	loi police projet payer intérêts
3 : journal septembre article suivant titre	journal septembre article suivant titre	journal septembre article suivant titre
4 : entier pourtant civils mots combien	entier pourtant mots combien vouloir	entier pourtant mots combien vouloir
5 : nombreux bois mètres maisons dernières	nombreux bois mètres maisons dernières	nombreux bois mètres maisons dernières
6 : bons divers assurer désormais chevaux	bons divers assurer désormais chevaux	bons divers assurer désormais chevaux
7 : elles trop chaque parce doute	elles trop chaque parce doute	elles trop chaque parce doute
8 : théâtre palais père revue comédie	théâtre palais père revue comédie	théâtre palais père revue comédie
9 : mer valdès marins côte appareil	mer valdès marins côte appareil	mer valdès marins côte appareil
10 : guerre deux dont leurs entre	guerre deux dont leurs entre	guerre deux dont leurs entre

Table 4: Most probable words for some topics from the historical French LDA model

4.3. Relevance

One problem with lift on its own is its tendency to give too much weight to words that are extremely rare in the corpus. Whilst these may be highly distinctive of the topic on the few occasions when they occur, they cannot be considered representative of the topic.

To deal with this, Sievert and Shirley [2014] also propose another metric, *relevance*. This is defined as a weighted average of the logarithms of likelihood and lift. The weighting must be either set by hand, on the basis of the desired importance of the distinctiveness of a term to a topic, or determined statistically.

Relevance of a term w in topic k is defined as:

$$rel(w|k) = \lambda * p(w|k) + (1 - \lambda) * p(w|k)/p(w)$$

where λ is a value from 0 to 1. Different values of λ give us different top relevant words. Notice that the last term is the definition for lift so if $\lambda = 0$ then relevance is just equal to the lift and if $\lambda = 1$ then the relevance is simply the probability of the word in the topic.

In Tables 3 and 4, we show the top words by probability, lift, and relevance (for $\lambda = 0.5$). In Section 6.2, we present a visualisation tool developed in [Sievert and Shirley, 2014] that allows the user to adjust λ interactively.

4.4. Bigram labels

Mei et al. [2007] experimented with using phrases to automatically generate topic labels, since their

previous work showed that human annotators prefer to use phrases when labelling topics manually. To extract phrases, they experimented with using chunking parsers to identify noun phrases and suitable bigrams from the documents in a corpus. They found that for one of their datasets, human evaluators clearly prefer bigrams over noun phrases while for another dataset, there is no clear difference. Both datasets are in English.

We experimented with using bigrams to label topics. For each topic k in a topic model, we do the following: (1) extract all bigrams for each document in the corpus where topic k is the topic with the highest probability; (2) score all bigrams from all the relevant documents using some relevance measure and output the top scoring bigrams.

4.4.1. Bigram scoring methods

Mei et al. [2007] introduced two relevance scoring methods that propose to measure the semantic similarity between the extracted bigrams and the respective topic.

Zero-order relevance score. This measures the relevance of a bigram b with a topic k using the topic-term distribution of k, β_k. Given bigram $b = w_1 w_2$, where w_n is word, the zero-order relevance score is:

$$score = log\frac{p(b|\beta_k)}{p(b)} = \sum_{0 \leq i \leq n} log\frac{p(w_i|\beta_k)}{p(w_i)}$$

The intuition behind this relevance score is that bigrams where both words have high probability for a topic should be relevant to that topic. It is called 'zero-order' because the measure is based only on the topic-term distribution without taking into account the corpus used to generate the topics.

First-order relevance score. This scoring function introduces a context C which is a collection of documents. C can be the same document collection used to train the topic model or a different collection. Given bigram b, the first-order relevance score is:

$$score(b|\beta_k) = \sum_w p(w|\beta_k) PMI(w, b|C) - D(\beta_k|C) + Bias(b|C)$$

where PMI is the pointwise mutual information score, $D(\beta_k|C)$ is the KL divergence between the context collection C and the topic. If C is the same corpus used to train the topic model, then the second term is zero. Lastly the bias term incorporates priors about the bigrams. This can also be set to zero if we do not wish to encode any prior information.

The intuition behind this scoring function is that the bigram must not only have a high relevance to the topic in terms of its topic-term distribution but it must also be relevant to a context. A context can be a collection of documents for which we want to generate our labels for.

We apply these scoring methods to the topics from the LDA models trained on the Reuters and historical French datasets. Tables 5 and 6 shows the top five bigrams of a few selected topics from the Reuters and historical French topics, respectively, ranked according to their zero-order and first-order relevance scores.

Topic words	Zero-order bigrams	First-order bigrams
1 : pakistan, northern, ireland, irish, lebanon, british	ireland loyalist, sinn fein, ira truce, ireland protestant, gerry adams	northern ireland, northern irish, catholic-based irish, irish guerrillas, irish republican
5 : wheat, prices, per, tonne, corn, grain,	toledo corn, toledo wheat, deliverable grades, chicago wheat, louis corn	toledo wheat, chicago wheat, wheat stocks, louis wheat, of wheat
7 : japan, yen, japanese, tokyo, yuan, shanghai	keizai shimbun, hiroshi mitsuzuka, nihon keizai, trillion yen, brokerage nomura	yen ($686,000), yen ($973,000), yen ($331,000), japan begins, 115 yen
10 : german, french, dollar, france, marks, mark	french francs, german marks, swiss francs, world currencies, one sterling	german marks, french francs, 1.7943/53 german, 1.6954/59 german, 1.5881/86 german
15 : oil, crude, production, bpd, barrels, per	bpd refinery, bpd crude, crude oil, meat bone, scalded edible	crude oil, oil supply, oil imports, oil stocks, oil industry

Table 5: Top bigrams from five topics of the Reuters topic model ranked by their zero-order and first-order relevance scores.

Topic words	Zero-order bigrams	First-order bigrams
1: rue saint gare mme train ami	dernière gare, gare de, denys cochin, un hôtel, hôtel du	5 rue, rue gustave-courbet, 77 rue, rue vavin, rue internationalee
8: théâtre palais père revue comédie mari	théâtre antoine, porte saint-martin, théâtre cluny, même spectacle, spectacle qu'en	théâtre antoine, théâtre sarah, théâtre réjane, théâtre michel, théâtre réjane
18: feu attaques gauche journée offensive obus	tranchée ennemie, démonstration offensive, aile droite, feu intense, offensive de	feu d'artillerie, feu d'arfillerie, feu intense, rive gauche, tranchée ennemie
26: munitions parti etats continue unis centre	continue argonne, parti socialiste, des munitions, munitions de, munitions d'artillerie	parti socialiste, munitions d'artillerie, continue argonne, etats doivent, parti socialiste
30: allemands armée troupes vers octobre bulgares	armée bulgare, armée turque, armée serbe, 9 octobre, octobre l'armée	allemands feraient, armée turque, allemands ont, troisième armée, grosse armée

Table 6: Top bigrams from five topics of the historical French topic model ranked by their zero-order and first-order relevance scores.

From a manual inspection of the Reuters bigrams, we do not see significant improvements of first-order bigrams over zero-order bigrams. One might say that in Topics 7 and 10, zero-order bigrams are more comprehensible since the first-order ones includes monetary amounts that do not lead to a better understanding of the topic. In Topic 15, however, first-order bigrams seem more relevant than the zero-order ones.

In the historical French bigrams, however, we see the advantages of first-order bigrams over the zero-order. The first-order bigrams are less likely to have common words such as 'un' or 'de' because the scoring metric includes a term for PMI. PMI scores bigrams whose words are more likely to be seen together than separately.

4.4.2. Finding high-coverage labels with embeddings

One issue with the zero-order and first-order scores is that they favour bigrams where both words have high probability in the topic. This is good for finding relevant bigrams but we not only want relevant labels, we also want labels that covers as many aspects of a topic as possible. For instance Topic 1 in Table 7 is not only about Ireland, it is also about conflicts in other parts of the world but the top bigrams are specific to events in Ireland. One way we can encourage high-coverage labels is to find labels that are more general or "stereotypical" of a topic.

We developed a method that uses clusters bigrams and selects the centroid bigrams with the idea that the centroid are more stereotypical bigrams. Our method works as follows: first, we compute bigram embeddings by taking the mean of the word embeddings in a bigram. Then with these bigram embeddings, we use a clustering method to cluster the embeddings and select the bigrams that are closest to the centre of each cluster. In our experiments, we take the top 100 bigrams of the zero-order bigrams, compute bigram embeddings using word embeddings trained on the Reuters dataset. Then we use k-means clustering (with $k = 5$) to cluster the bigram embeddings and take the centroid of each the five clusters.

The results of this experiment are shown in Table 7. In Topic 1, compared with the zero-order bigrams, the centroid bigrams have more general phrases that encompasses more of the topic. For Topics 5, 10, and 15, however, this method is less successful.

4.5. Labelling of dynamic topics

Dynamic topic model [DTM, Blei and Lafferty, 2006] captures topic evolution in the corpus. This means that the probability distribution of a topic changes slightly from one time slice to the next and therefore the top words (according to some metric) also vary. The top words of a topic in the first time slice might look quite different from the top words in the last time slice. Generating textual topic descriptions becomes more complicated in this case, especially as some topics that might undergo significant changes in their topic-term distributions due to, for instance, significant events reported in the news.

We experimented with getting a kind of *mean top topic words* by taking the mean probability distribution of a topic over all time slices and computing the top words from this mean distribution.

Table 8 shows the most probable words on a topic about church and clergy in Finland over six time

Topic words	Zero-order bigrams	Centroid bigrams
1 pakistan, northern, ireland, irish, lebanon, british	ireland loyalist, sinn fein, ira truce, ireland protestant, gerry adams	ireland security, substantive peace, guerrilla arms, protestant politicians, bomb attacks
5 wheat, prices, per, tonne, corn, grain,	toledo corn toledo wheat, deliverable grades, chicago wheat, louis corn	chicago, deliverable, 260 ccc, winter 5,306, grades 5,306, dark northern
7 japan, yen, japanese, tokyo, yuan, shanghai	keizai shimbun, hiroshi mitsuzuka, nihon keizai, trillion yen, brokerage nomura	affiliate nomura, june koike, to racketeers, mof spokesman, nomura executives
10 german, french, dollar, france, marks, mark	french francs, german marks, swiss francs, world currencies, one sterling	francs 6.2795/15, francs 1711.5/3.0, marks 1.9023/28, french francs, francs 6.0350/70
15 oil, crude, production, bpd, barrels, per	bpd refinery, bpd crude, crude oil, meat bone, scalded edible	refinery was, bpd crude, 180,000 bpd, 305,420 crude, tallow, edible

Table 7: Top bigrams from five topics of the Reuters topic model ranked by their zero-order and first-order relevance scores.

	Finnish DTM - Topic 11
1854	kappalainen apul virka apulainen kirkkoherra
1855	kappalainen apul virka apulainen kirkkoherra
1856	kirkkoherra kappalainen virka apul apulainen
1870	virka opettaja kappalainen kirkkoherra turku
1871	virka kappalainen opettaja ylioppilas kirkkoherra
1872	virka kappalainen opettaja kirkkoherra turku
mean topic words	kirkkoherra virka kappalainen opettaja turku

Table 8: Most probable words of a dynamic topic on the church and clergy in Finland for three consecutive time slices (1854-1856) followed by another three consecutive time slices (1870-1872) and finally the mean most probable words.

slices (1854-1856 and 1870-1872) from the DTM trained on Finnish news and the mean most probable words for all 64 time slices.

In Table 9, we show an example of a topic that undergoes significant changes periodically. This topic is about the legislative assembly of Finland in the nineteenth-century, the Diet of Finland, that conducted a session every few years or so. When a session occurred, as in 1872 and 1877-78, *valiokunta* (committee) becomes a top word but drops in prominence when there is no session. The mean topic words include *sääty* (estate) and *kokous* (meeting) but do not include *valiokunta* (committee). These group of words describe the topic adequately but *valiokunta* is a significant term that would communicate the gist of the topic more effectively.

4.6. Multilingual topic labelling with deep learning

A more recent development in automatic topic labelling is using deep learning to directly generate labels. Alokaili et al. [2020] proposed a sequence-to-sequence model (seq2seq) trained on a synthetic

	Finnish DTM - Topic 16
1871	sääty kokous kysymys esitys valtio
1872	sääty mietintö **valiokunta** esitys mainita
1873	herra sääty kysymys mietintö mainita
1876	herra sääty kysymys mietintö puhuja
1877	sääty mietintö herra puhuja **valiokunta**
1878	sääty päätös **valiokunta** ehdotus mietintö
1879	sääty päätös ehdotus esitys **valiokunta**
1880	sääty päätös esitys keskustelu ehdotus
mean topic words	kokous kysymys ehdotus pitää sääty

Table 9: Tracking the significance of *valiokunta* in a dynamic topic about the Finnish legislative assembly. Years in bold are when a session of the legislative assembly took place.

dataset of Wikipedia articles and titles while Popa and Rebedea [2021] finetuned BART, a pretrained transformer-based language model [Lewis et al., 2020], with topic keywords and candidate labels from weak labellers to generate labels. However, there has been no work so far on coming up with multilingual topic labels. Generating labels in multiple languages allows users to compare topical trends across linguistic boundaries without having to align topics and to explore news collections by users who might not have the necessary linguistic skills to do otherwise.

To produce multilingual labels for news topics, we proposed an ontological mapping method that maps topics to concepts in a language-agnostic news ontology. These concepts have labels in multiple languages that we used as topic labels. We approached ontology mapping as a multilabel classification task where a topic can be classified as belonging to multiple concepts. This work has been accepted to the 44th European Conference on Information Retrieval (ECIR 2022) to be held in Stavanger, Norway on 10-14 April 2022. The manuscript is attached to this deliverable.

4.6.1. Models

Ontology Mapping. The classifier takes as an input a sequence $X = (x_1, \ldots, x_n)$ of the n top terms of a topic, and predicts $P(c_i|X)$, the probabilities for each ontology concept $c_i \in C$. The topic labels are obtained from the distribution $P(c_i|X)$ as follows: First, a list of label candidates was obtained by considering all c_i such that $P(c_i|X) > t$, where t is the classification threshold. Then, we propagated the predicted concepts to the top of the ontology. For instance, if a topic is classified as belonging to concept 01005000:CINEMA, it also belongs to concept 01000000:ARTS, CULTURE AND ENTERTAINMENT, the parent of 01005000:CINEMA. Lastly, we obtained the topic labels by taking the most frequent concepts among the candidates and taking the labels of these concepts in the preferred language.

To compute the probabilities $P(c_i|X)$, we encoded the top terms $(x_1, \ldots, x_n)$ using SBERT [Reimers and Gurevych, 2019][3] and passed this representation to a classifier composed of two fully-connected layers with a ReLU non-linearity and a softmax activation. We set the classification threshold t to 0.03 as determined by the validation set. We refer to this as the **ontology** model. We illustrate this model in Figure 2.

[3]We use the multilingual model *distiluse-base-multilingual-cased*.

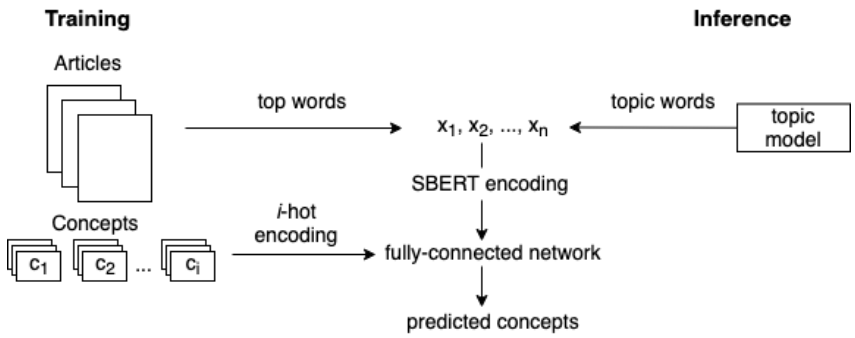


Figure 2: News concepts prediction pipeline.

Comparisons to State-of-the-art. We also investigated how our ontology mapping method compares to methods that directly generate topic labels. Alokaili et al. [2020] used an RNN-based encoder-decoder architecture with attention as a seq2seq model while Popa and Rebedea [2021] finetuned a pretrained BART model. Both methods have reported state-of-the-art results on English topics from multiple domains.

We implemented a RNN seq2seq model using the same hyperparameters as Alokaili et al. [2020]: 300 dimensions for the embedding layer and a hidden dimension of 200. We refer to this as the **rnn** model. We also implemented a slightly modified model where we replaced RNN with transformers, which has yielded state-of-the-art results in many NLP tasks. We used the hyperparameters from the original transformers model [Vaswani et al., 2017]: 6 layers for the encoder and decoder with 8 attention heads and a embedding dimension of 512. We refer to this as the **transformer** model.

Instead of BART which is trained only on English, we finetuned a multilingual version, mBART [Liu et al., 2020], and set the source and target languages to Finnish. We finetuned mBART-25 from HuggingFace[4] for 5 epochs. We used the AdamW optimizer with weight decay set to 0.01. We refer to this as the **mbart** model [5]. For consistency, all the models except mbart were trained using Adam optimizer for 30 epochs with early stopping based on the validation loss.

4.6.2. Datasets

News Ontology. We used the IPTC Subject Codes as our news ontology.[6] This is a language-agnostic ontology designed to organise news content. Labels for concepts are available in multiple languages - in this work we focused specifically on Finnish and English. This ontology has three levels with 17 high-level concepts, 166 mid-level concepts and 1221 fine-grained concepts. Mid-level concepts have exactly one parent and multiple children.

Training Data. We used news articles from 2017 of the Finnish News Agency dataset [STT, 2019, STT et al., 2020] which have been tagged with IPTC concepts and lemmatized with the Turku neural parser [Kanerva et al., 2018]. Following the distant-supervision approach in Alokaili et al. [2020], we constructed a dataset where the top n words of an article are treated as input $X = (x_1, \ldots, x_n)$ and

[4] https://huggingface.co/facebook/mbart-large-cc25
[5] While the mBART encoder is in a multilingual space, it cannot be used directly for cross-lingual language generation [Maurya et al., 2021].
[6] https://cv.iptc.org/newscodes/subjectcode/

the tagged concepts are the target C; an article can be mapped to multiple concepts. Top words can either be the top 30 scoring words by tf-idf (**tfidf** dataset) or the first 30 unique content words in the article (**sent** dataset). We trained all models on both datasets. For each dataset, we have 385,803 article-concept pairs which we split 80/10/10 into train, validation and test sets.

Test Data. For Finnish topics, we trained an LDA model for 100 topics on the articles from 2018 of the Finnish news dataset and selected 30 topics with high topic coherence for evaluation. We also checked that the topics were diverse enough so that they cover a broad range of subjects.

To obtain gold standard labels for these topics, we recruited three fluent Finnish speakers to provide labels for each of the selected topics. For each topic, the annotators received the top 20 words and three articles closely associated with the topic. We provided the following instructions to the annotators:

Given the words associated with a topic, provide labels (in Finnish) for that topic. There are 30 topics in all. You can propose as many labels as you want, around 1 to 3 labels is a good number. We encourage concise labels (maybe 1-3 words) but the specificity of the labels is up to you. If you want to know more about a topic, we also provide some articles that are closely related to the topic. These articles are from 2018.

We reviewed the given labels to make sure the annotators understood the task and the labels are relevant to the topic. We used all unique labels as our gold standard, which resulted in seven labels for each topic on average. While previous studies on topic labelling mainly relied on having humans evaluate the labels outputted by their methods, we opted to have annotators *provide* labels instead because this would give us an insight into how someone would interpret a topic[7]. During inference, the input X were the top 30 words for each topic.

To test our model in a cross-lingual zero-shot setting, we used the English news topics and gold standard labels from the NETL dataset [Bhatia et al., 2016]. These gold labels were obtained by generating candidate labels from Wikipedia titles and asking humans to evaluate the labels on a scale of 0-3. This dataset has 59 news topics with 19 associated labels but we only took as gold labels those that have a mean rating of at least 2.0, giving us 330 topic-label pairs. We used default topic labels—top five terms of each topic— as the baselines.

4.6.3. Results and Discussion

We used BERTScore [Zhang et al., 2019] to evaluate the labels generated by the models with regards to the gold standard labels. BERTScore finds optimal correspondences between gold standard tokens and generated tokens and from these correspondences, recall, precision, and F-score are computed.

We show the average BERTScores for the Finnish news topics at the top of Table 10. All models outperformed the baseline by a large margin which shows that labels to ontology concepts are more aligned with human-preferred labels than the top topic words. The rnn-tfidf model obtained the best scores followed by ontology-sent. The transformer-sent and mbart-sent models also obtained comparable results. We did not see a significant difference in performance between training on the tfidf or sent datasets. In Table 11 (top), we show an example of the labels generated by the models and the gold standard labels.

[7] Volunteers are compensated for their efforts. We limited our test data to 30 topics due to budget constraints.

	PREC	REC	F-SCORE
Finnish news			
baseline: top 5 terms	*89.47*	*88.08*	*88.49*
ontology-tfidf	94.54	95.42	94.95
ontology-sent	95.18	**95.96**	95.54
mbart-tfidf	93.99	94.56	94.19
mbart-sent	94.02	95.04	94.51
rnn-tfidf	**96.15**	95.61	**95.75**
rnn-sent	95.1	94.63	94.71
transformer-tfidf	94.26	94.42	94.30
transformer-sent	95.45	94.73	94.98
English news			
baseline: top 5 terms	***98.17***	***96.58***	***97.32***
ontology-tfidf	97.00	95.25	96.04
ontlogy-sent	97.18	95.43	96.21

Table 10: Averaged BERTScores between labels generated by the models and the gold standard labels for Finnish and English news topics.

It can be seen that the baseline consists mostly of proper names—Räikkönen, Bottas, Hamilton[8]—that contradicts the main idea of the topic modelling, which should represent collection themes rather than specific facts. All models gave sufficiently suitable labels, focusing on motor sports. However only the ontology-sent model was able to get Formula 1 as one of its labels. Moreover, the English labels were taken directly from the ontology model and not from manual translations.

We also demonstrated the ability of the ontology models to label topics in a language it has not seen during training by testing it on English news topics from the NETL dataset [Bhatia et al., 2016]. We encoded the topic words with SBERT and passed them to the trained ontology models, which have been trained only on Finnish articles. This dataset was also used in Alokaili et al. [2020] for testing but our results are not comparable since they presented the scores for topics from all domains while we only used the news topics. The results are shown at the bottom of Table 10. Although the ontology models did not outperform the baseline, they were still able to generate English labels that were very close to the gold labels considering that it has only been trained on Finnish. From the example in Table 11 (bottom), we also observed that the gold labels are overly specific, suggesting names of directors as labels when the topic is about the film industry in general. We believe this is due to the procedure used to obtain the gold labels, where the annotators were asked to *rate* labels rather than propose their own.

4.6.4. Experiments with historical topics

We experimented with applying the deep learning methods described above to topics extracted from historical news collections. We used articles from *Uusi Suometar*, a Finnish language newspaper published in the nineteenth-century.[9] Unlike the modern news dataset, these articles were not tagged with IPTC tags.

[8] Incorrect spelling of 'hamilton' is because of lemmatisation.

[9] Available from the National Library of Finland: https://digi.kansalliskirjasto.fi/sanomalehti/titles/1457-4721?display=THUMB&year=1918

	Finnish topic
Topic	räikkönen, bottas, ajaa *(to drive)*, hamilto, mercedes
Gold	formula, formulat, formula 1, f1, formula-auto, aika-ajot *(time trial)*, moottoriurheilu *(motor sport)*
rnn-tfidf	autourheilu *(auto sport)*, urheilutapahtumat *(sports event)*, mm-kisat *(world championship)*, urheilu *(sport)*, urheilijat *(athletes)*
transformer-sent	urheilutapahtumat *(sports event)*, mm-kisat *(world championship)*, urheilu *(sport)*, autourheilu *(auto sport)*, kansainväliset *(international)*
mbart-sent	autourheilu moottoriurheilu, urheilutapahtumat, mm-kisat , urheilijat pelaajat, urheilu
ontology-sent	ID: 15000000, fi: urheilu, en: sport; ID: 15039000, fi: autourheilu moottoriurheilu, en: motor racing; ID: 15073000, fi: urheilutapahtumat, en: sports event; ID: 15039001, fi: formula 1, en: formula one; ID: 15073026, fi: mm-kisat, en: world championship
	English topic
Topic	film, movie star, director, hollywood, actor, minute, direct, story, witch
Gold	fantasy film, film adaptation, quentin tarantino, a movie, martin scorsese, film director, film
ontology-sent	ID: 01005001, en: film festival, fi: elokuvajuhlat; ID: 04010003, en: cinema industry, fi: elokuvateollisuus; ID: 08000000, en: human interest, fi: human interest; ID: 01022000, en: culture (general), fi: kulttuuri yleistä; ID: 04010000, en: media, fi: mediatalous

Table 11: Generated labels for selected topics. Finnish labels are manually translated except for ontology-sent. For ontology-sent, we provide the concept ID and the corresponding Finnish and English labels.

Training data In addition to training models on the news articles from the Finnish News Agency, we also trained models on a dataset of Finnish Wikipedia articles. We hypothesized that since Wikipedia had articles on Finnish history, it might be more informative for historical topics than a modern news dataset. Instead of using IPTC tags as in the news dataset, we used the article titles and categories as prospective labels. As with the modern news dataset, our training data was composed of article-title or article-category pairs.

Test data We trained an LDA model for 50 topics on *Uusi Suometar* articles from 1860-1879 and select 20 topics with high coherence scores for evaluation. To obtain gold standard labels, we asked two historians trained in nineteenth-century Finnish history to label the topics. We obtained, on average, three labels per topic with each label being one token long.

Historical News			
	PREC	REC	F-SCORE
baseline: top 5 terms	***88.1***	***86.19***	***87.09***
ontology-sent	85.44	85.06	85.16
mbart-news-tfidf	83.00	84.02	83.43
mbart-news-sent	85.09	85.16	85.04
mbart-wiki-tfidf	80.93	81.67	81.15
mbart-wiki-sent	79.52	81.34	80.24

Table 12: BERTScore between generated and gold labels for historical topics.

Results and Discussion We present evaluation results in Table 12. Unlike the modern news topics, automatically generated labels did not perform as well with historical topics. The baseline labels correlated more with the labels provided by historians than those generated by any of our models. The models were trained on modern data but resources such as Wikipedia contain information about historical events and concepts and we had hoped that the controlled vocabulary of the IPTC taxonomy might be be broad enough in order to be able to also describe topics from the nineteenth-century.

We present some topics and their generated labels in Table 13. The first topic is about shipping and navigation. The labels provided by the historians were quite concise and contained the top word of the topic 'meri' (sea). The labels generated by the model trained on Wikipedia were mostly off-topic while the labels from the models trained on modern news were about maritime travel as a leisure activity than as a profession since concepts related to shipping (merenkulku) and navigation do not exist in the IPTC taxonomy.

Another issue we noticed is the problem of anachronistic labels as shown in the second example—a topic on foreign politics. The models generated names of organisations and institutions such as the European Commission and the Finnish parliament (*eduskunta*) but these entities did not exist when these articles were published (1860-1879).

Another aspect to consider when it comes to historical topics is that there are at least three time periods and their respective vocabularies to take into account. First, there is the level of reception: how a certain period was seen and described in the period following it. Second, the level of questioning that happens in the present and thus refers to our current vocabulary. Lastly, the time level which is a description of the situation or the object in the context of the time (contemporary accounts) and the respective vocabulary of the time.

4.6.5. Conclusions

We proposed a straightforward ontology mapping method for producing multilingual labels for modern news topics. We cast ontology mapping as a multilabel classification task, represented topics as contextualised cross-lingual embeddings with SBERT and classified them into concepts from a language-agnostic news ontology where concepts have labels in multiple languages. Our method performed on par with state-of-the-art topic label generation methods, produced multilingual labels, and worked on multiple languages without additional training. We also showed that labels of ontology concepts correlated highly with labels preferred by humans.

	Historical topics
Topic 1	meri *(sea)*, saari *(island)*, ranta *(beach)*, laiva *(ship)*, laima *(OCR-laiva)*, sataa *(rain)*, englanti *(England)*, kapteeni *(captain)*
Gold	merimatkat *(sea voyages)*, merenkulku *(shipping)*
ontology-sent	ID: 04000000, fi: talous, en: economy, business and finance; ID: 8000000, human interest; ID: 6000000, fi: ympäristö, en: environmental issue; ID: 10000000, fi: vapaa-aika, en: lifestyle and leisure; ID: 11000000 fi: politiikka, en: politics
mbart-wiki-tfidf	amerikkalainen geologia *(American geology)*, laivatyypit *(types of ship)*, meri *(sea)*, aida cruises, amerikkalainen meri *(American sea)*
mbart-news-sent	vesiliikenne *(water transport)*, vapaa-aika *(leisure)*, lomamatkailu turismi *(holiday tourism)*, liikenne *(transport)*, vesiliikenneonnettomuudet *(waterway accidents)*
Topic 2	hallitus *(government)*, ranska *(France)*, ministeri *(minister)*, italia *(Italy)*, ulkomaa *(abroad/foreign)*, espanja *(Spain)*, puolue *(party)*, kuningas *(king)*
Gold	kansainvälinen politiikka *(international politics)*, ulkomaat *(foreign countries)*
ontology-sent	ID: 11000000 fi: politiikka, en: politics; ID: 11006000, fi: julkinen hallinto, en: government; ID: 02000000, fi: laki oikeus rikokset, en: crime, law, and justice; ID: 14000000, fi: sosiaalikysymykset, en: social issue; ID: 11009000, fi: eduskunta, en: parliament
mbart-wiki-tfidf	espanjalaiset puolueet *(Spanish parties)*, espanjalaiset eu-komissaarit *(Spanish EU commissioners)*, ranskan politiikka *(French politics)*, italialaiset puolueet *(Italian parties)*, politiikan käsitteet *(political concepts)*
mbart-news-sent	eduskuntavaalit presidentinvaalit *(parliamentary and presidential elections)*, julkinen hallinto *(public administration)*, politiikka *(politics)*, valtionpäämiehet *(heads of state)*, ministerit *(ministers)*

Table 13: Generated labels for some topics from historical Finnish news.

We applied the label generation and ontology mapping methods to topics from a historical Finnish newspaper and asked historians to provide gold standard labels. Unlike modern news, our methods did not outperform the baseline labels (top topic words by probability). We hypothesise that this is because our training datasets (modern news articles and Wikipedia) were written for a modern context and we did not have a large enough training dataset that is annotated for a specific historical period (by annotations, we refer to the IPTC concept tags in the news dataset or the titles and categories in the Wikipedia dataset). In future work, we could improve this by selecting only titles of Wikipedia articles

categorised under history.

5. Textual topic summaries

An alternative approach to represent topics of a topic model (TM) trained over a document collection is to generate a summary of the documents with high topic probabilities (according to the *document-topic distributions* $p(k|d)$ estimated from the TM, see Section 2). The result is a textual output consisting of *sentences* (in contrast with *labels*, as obtained in Section 4).

The textual summary for a topic k is constructed in two steps:

- document selection (to identify the documents in the collection under exam that are relevant for the topic k);
- multi-document summarisation (to obtain a summary of the selected documents).

In our experiments, we considered the TM trained with the Reuters and NewsEye French datasets of Section 3. In the following, we present the document selection strategy adopted in the experiments (Section 5.1), and introduce different summarisation methods (Section 5.2). Results and a comparison of the textual topic summaries obtained are reported in Section 5.3.

5.1. Document Selection

Using the *document-topic distributions* estimated from the topic model (TM) trained on a collection of documents $\mathcal{C}$, it is possible to compute the probability $p(k|d)$ of a topic k given a document d. The probability can be used to identify a selection $\mathcal{C}_k^n$ of n documents that are relevant to the topic. This can be done in two ways:

- *Split and sort*, consisting in the following steps:
  1. associate each document $d \in \mathcal{C}$ to a single topic k_d by $k_d = \underset{k}{\operatorname{argmax}}\, p(k|d)$;
  2. for each topic k, consider the set of documents $\mathcal{C}_k = \{d \in \mathcal{C} : k_d = k\}$;
  3. sort the documents in $\mathcal{C}_k$ by their probabilities $p(k|d)$ and select the n elements with highest probability (or all the elements if $|\mathcal{C}_k| \leq n$) to obtain $\mathcal{C}_k^n$.
- *Sort by topic*: for each topic k, $\mathcal{C}_k^n$ is built by taking the n documents in the collection $\mathcal{C}$ with highest $p(k|d)$.

We initially considered both approaches and observed how the *Split and sort* method can lead to unbalanced topic collections $\mathcal{C}_k$, with dimensions spanning over several orders of magnitude, from $\mathcal{O}(1)$ to $\mathcal{O}(|\mathcal{C}|)$. For this reason, we stick to the *Sort by topic* strategy. It has to be noted that, by using the latter, some of the documents in the topic collection $\mathcal{C}_k^n$ might not have k as most prominent topic. In any case, the documents in the collection will contain – up to a certain degree – the topic k.

5.2. Summarisation Methods

After document selection, we produce textual topic summaries for each topic k, using the documents $\mathcal{C}_k^n$. For this purpose, we consider several *multi-document summarisation methods*. Text summarisation methods are generally classified in three main classes: *extractive* (given a document - or a collection of

documents - the algorithm *extracts* the most relevant sentences), *abstractive* methods (generates short summaries capturing salient ideas from the input text) and *compressive* (each sentence of the input text is reduced in length preserving the initial information).

Our experiments are limited to *extractive* methods, as accuracy and explainability are our main concerns, and these models are generally more transparent. Moreover, abstractive and compressive techniques work by paraphrasing or shortening the input text, so that the original meaning is not always preserved. By making use of extractive techniques, we ensure that content is preserved at the sentence level.

We build on the summarisation methods used in Cano Basave et al. [2014], where the authors show that they can be successfully used for topic labelling. Below, we list the methods considered in our experiments. Note that these techniques do not aim at producing coherent summaries: the output text is indeed a set of disjointed sentences that are representative of the input documents.

5.2.1. First Sentence

First, we consider a very simple algorithm. Given a topic collection $\mathcal{C}_k^n$, the documents are sorted by their topic probability $p(k|d)$. All the first sentences which contains at least one of the top 10 words for the topic k are concatenated to generate the summary. Other sentences are considered with the same ordering until the desired summary length is reached.

5.2.2. Sum Basic

Sum Basic (SB) [Nenkova and Vanderwende, 2005] is a summarisation algorithm based on word frequency. Given a collection of documents, it selects sentences as follows:

- Word probabilities are initially computed over the input documents;
- At each step:
  - The sentence with the highest average word probability is selected (among all the sentences in the input documents) and is added to the summary;
  - To ensures a certain degree of textual variability in the final summary, word probabilities for the tokens of the selected sentence are scaled down by squaring them.

In our implementation, the initial word probabilities are not computed over the collection $\mathcal{C}_k^n$, but we make use of the *topic-word distribution* from the TM. The idea behind is that probabilities from the TM captures topics that might not be frequent when considering only the top n documents from $\mathcal{C}_k$.

5.2.3. Hybrid TFIDF

This is a variation of the Sum Basic algorithm illustrated above, where initial word probabilities are replaced by TFIDF scores [Cano Basave et al., 2014]. For each word that is present in a collection of

documents for the topic k, its Term Frequency (TF) is computed as

$$\mathrm{TF}(w,k) = \frac{f_{w,\mathcal{C}_k^n}}{\sum_{w' \in \mathcal{C}_k^n} f_{w',\mathcal{C}_k^n}}, \tag{1}$$

where $f_{w,\mathcal{C}_k^n}$ is the word frequency in the topic collection $\mathcal{C}_k^n$. Inverse Document Frequency (IDF) is computed by

$$\mathrm{IDF}(w,k) = \log \frac{|\hat{\mathcal{C}}_k^n| + 1}{|\{d \in \hat{\mathcal{C}}_k^n : w \in d\}| + 1} \tag{2}$$

where $\hat{\mathcal{C}}_k^n = \{d \in \mathcal{C}_t^n \, \forall \mathcal{C}_t^n : t \neq k\}$. Finally, TFIDF scores are given by the product of the two terms above,

$$\mathrm{TFIDF}(w,k) = \mathrm{TF}(w,k) \cdot \mathrm{IDF}(w,k). \tag{3}$$

Note that there are several option to define derive TFIDF scores, e.g. by computing the TF over a single document, or by taking the IDF as the total number of documents in the whole corpus containing certain words. The rationale for our choices is that:

- When computing $\mathrm{TF}(w,k)$, counting word frequencies over the topic collection $\mathcal{C}_k^n$ – instead of over the single document – rewards terms that are more frequent within the topic under analysis;
- In the denominator of Equation 2, documents associated to the topic k are not counted, so that words appearing in all documents of the analysed topic are not penalised.

We will refer to this variant of the SB algorithm as *Hybrid TFIDF*, even if our definition differs from the one introduced in Cano Basave et al. [2014].

5.2.4. Text Rank

Text Rank [Mihalcea and Tarau, 2004] is a graph-based summarisation method. Each vertex of the graph represents a word. Vertex weights are computed recursively from the global graph structure and are associated with word relevance. Text Rank is based on Page Rank, an algorithm initially introduced for web searches [Brin and Page, 1998].

We consider two flavours of Text Rank. One variant computes the average relevance for all sentences in the topic collection, and picks the most relevant sentence at each iteration step until the desired summary length is reached. The second variant works in the same way, but the most relevant sentences are discarded if they are too similar to sentences that are already in the summary. Sentence similarity is computed by using Word2Vec embeddings [Mikolov et al., 2013] trained over the entire Reuters corpus. In the following, we refer to the former as *Text Rank Most Relevant* (TR MR) and to the latter as *Text Rank with Similarity* (TR Sim).

5.3. Experiments

Below, we discuss the the experimental setup adopted to test the summarisation techniques introduced in the previous section with the Reuters TM (see Section 3.3). We also report example summaries obtained for one topic of the historical French news dataset TM of Section 3.2.

In the document selection step (crf. Section 5.1), we consider the top 50 documents ($n = 50$) for each topic, then run the summarisation algorithms over the topic collections $\mathcal{C}_{k_i}^{50}$ (where $i = 1, \ldots, 50$ represent

the topic label). For Sum Basic and Hybrid TFIDF methods, stop words and tokens of length 2 or less are removed from the text before computing initial word probabilities. All the considered summarisation techniques work by selecting sentences from the topic collection $\mathcal{C}_{k_i}^{50}$, we limit the number of sentence by stopping the summary generation when the result contains 100 words or more. As such, all obtained summaries have a length bigger or equal to 100 terms.

In Tables 14 and 15 we report, respectively, the summaries obtained for Topic 1 of the Reuters TM and Topic 26 of the historical French TM. For a given topic, the summaries generated can overlap, as some of the algorithms rely on common features to select the most relevant phrases. In both tables, sentences that are selected by multiple algorithms are highlighted in colours.

Comparing the Reuters summaries of Table 14 with the TM top words and the bigram labels obtained for the same topic (reported in the first row of Tables 5 and 7), we note that there is no mention of "Pakistan", but the summaries are consistent with all the other top words and bigram labels. In this specific case, the Sum Basic summary can provide a better topic coverage compared to bigram labels, as it also mentions civil war in Lebanon, that is not covered by bigram representations, despite it being – according to the TM top words – a prominent subject within the topic.

Summaries generated for the historical French TM have a similar number of overlapping sentences. The SB method and its TFIDF variant seem to provide a better coverage when compared to the bigram topic representations given in Table 6: SB and TFIDF summaries contain references to the Socialist Party, which appears both in the zero-order and first-order bigram representations and is absent in the summaries generated by TR-based methods.

Our observations do not necessarily indicate that SB-based summarisation algorithms give better results. Both bigram representations and SB algorithms rely primarily on topic word probabilities (differently from TR-based methods – where word importance is based on graph weights). This could explain why they provide summaries that seem more consistent with top words or bigram labels. In conclusion, we cannot gauge quality of the different summaries without performing a human evaluation of these representations.

To inspect similarities between the different generation methods, we perform pairwise comparisons by computing BLEU scores [Papineni et al., 2002]. The results, reported in Table 16, are derived by using an open-source implementation of the BLEU score with default parameters.[10] The topic summaries obtained for the Reuters TM from each summarisation technique are used both as text source and reference. The obtained scores are low, denoting that the different methods select different phrases as most relevant. As expected, the quantitative comparison shows that the two Text Rank methods are the most similar, with a BLEU of 0.39, and are quite different from the Sum Basic-based techniques (with BLEU scores always lower than 0.1, with a lowest value of about 0.04 when comparing SB and TR-MR). Sum Basic and Hybrid TFIDF achieve BLEU scores of about 0.2 when comparing the generated summaries. Hybrid TFIDF is the closest to the First Sentence method in terms of BLEU scores.

6. Topic model visualisation

One of the key conclusions we have arrived at from our investigation of representation methods and our discussions with DH researchers is that, whilst textual representations of topics are essential for

[10] https://github.com/tuetschek/e2e-metrics

Method	Summary of Topic 1
SB	The loyalists want northern Ireland to remain British. Britain and Sinn Fein have both denied an Irish newspaper report suggesting that the IRA was operating an unofficial ceasefire to get Sinn Fein into the talks. Security sources in Lebanon said on Thursday night that one guerrilla was wounded when two Israeli helicopters fired five rockets at suspected Hezbollah targets in south Lebanon. Loyalist politicians linked to the protestant guerrillas urged them to heed a yearning for peace and maintain the truce. Britain says that the province will remain under London rule while a majority in the region support that.
TFIDF	Britain and Ireland have said Sinn Fein will not be allowed into talks until the IRA declares a full ceasefire. The British and Irish governments insist on a new truce before all-party talks on British rule in northern Ireland can begin. Loyalist politicians linked to the protestant guerrillas urged them to heed a yearning for peace and maintain the truce. But it stayed at the talks when the democratic unionist party and UK unionist party walked out last week in protest at what they called an appeasement of IRA terrorism. Police said no bomb warning was given by any of the province's rival guerrilla groups.
TR MR	And it is certainly not the right way to make progress, Major said Sinn Fein has been excluded from multi-party peace talks that started in Belfast in June because of the IRA's refusal to reinstate the ceasefire it broke in February last year citing what it called British intransigence in the peace process. Irish Prime Minister John Bruton accused the IRA on Thursday of trying to provoke its pro-British Unionist rivals into breaking their truce and said Republicans would be banned from Irish peace talks until they end violence.
TR Sim	And it is certainly not the right way to make progress, Major said Sinn Fein has been excluded from multi-party peace talks that started in Belfast in June because of the IRA's refusal to reinstate the ceasefire it broke in February last year citing what it called British intransigence in the peace process. Police said no bomb warning was given by any of the province's rival guerrilla groups. The Loyalist paramilitaries have maintained their ceasefire. It was made safe by army experts. The driver escaped on foot.

Table 14: Example summaries generated for one example topic of the Reuters topic model by applying the summarisation methods of Section 5.2. Sentences that are picked by multiple algorithms are reported with the same colour. Top bigrams for the topic considered in this example are reported in Tables 5 and 7 (Topic 1).

Method	Summary of Topic 26
SB	Les convois de munitions n'ont pas bougé. C'est eux qui ont provoqué la déclaration de la commission permanente du parti socialiste. La note ajoute que les bonnes relations des deux etats doivent reposer sur le respect du droit et de l'humanité, formule qui rend conditionnelles les bonnes relations. La bataille de champagne continue... plus au nord. Cette partie centrale, comme un vulgaire centre politique, se décompose en deux parties: centre droit et centre gauche. Les socialistes rétormistes se sont abstenus. Tantôt de petits drapeaux tricolores, tantôt des tas de châtaignes.
TFIDF	La bataille de champagne continue... plus au nord. Il y a des socialistes que cette guerre embarrasse. Sembat, Thomas se sont conduits devant le congrès de leur parti en patriotes. Souhaitons-le, car uskub entre les mains serbes ce serait la possibilité de déblayer la voie ferrée jusqu' à nich et d'assurer les communications des alliés. Ce col donne passage à la route de veles à monastir par Prilep, au travers du massif de Babuna. Tantôt de petits drapeaux tricolores, tantôt des tas de châtaignes. Ces deux représentants de la minorité socialiste française se prononcent pour la reprise immédiate des rapports , sans condition préalable .
TR MR	Depuis un certain temps, et surtout depuis la bataille de Loos où l' on commit les mêmes erreurs tactiques qu'à Neuve - Chapelle, dans les cercles officiels on javait agité divers projets et voici que le 20 novembre dernier, pour la première fois en Angleterre, un journal, l' Observer, demande que des officiers français commandent directement des soldats du roi. Le directeur de L' Œuvre et les collaborateurs groupés autour de lui, depuis tantôt douze ans, n' ont jamais eu d' autre but; L' avantage serait double; Autriche.
TR Sim	Depuis un certain temps, et surtout depuis la bataille de Loos où l' on commit les mêmes erreurs tactiques qu'à Neuve - Chapelle, dans les cercles officiels on javait agité divers projets et voici que le 20 novembre dernier, pour la première fois en Angleterre, un journal, l' Observer, demande que des officiers français commandent directement des soldats du roi. C' est pour être « utiles » qu' ils ont mené — et avec quelle vigueur et avec quelle ténacité! cial change constamment d' aspect. Autriche. Exemples. La Chambre.

Table 15: Example summaries generated for one example topic of the historical French topic model by applying the summarisation methods of Section 5.2. Sentences that are picked by multiple algorithms are reported with the same colour. Top bigrams for the topic considered in this example are reported in Table 6 (Topic 26).

Method \ Reference	FS	SB	TFIDF	TR Sim	TR MR
FS	1.000	0.094	0.143	0.106	0.102
SB	0.091	1.000	0.233	0.056	0.037
TFIDF	0.143	0.238	1.000	0.093	0.089
TR Sim	0.102	0.056	0.091	1.000	0.394
TR MR	0.100	0.038	0.088	0.394	1.000

Table 16: BLEU scores obtained by comparing summaries generated for the 50 topics of the Reuters TM with the summarisation methods of Section 5.2.

some purposes in NewsEye, other descriptive forms of representation of topics can also be valuable for understanding what is captured by the model.

In this section, we present the visualisation methods we have explored to visualise TMs and individual topics. We make use of some well-known visualisation techniques such as word clouds and LDAVis for LDA models and explore different ways of visualising different aspects of dynamic topic models.

Visualisations can show interesting properties of topics that might not be apparent from a textual description. Interactive visualisations also allow the user to explore different topic representations by changing the parameters themselves. Moreover, when it comes to more complex TMs such as dynamic topic models, plots and graphs can convey information in a concise manner that might be more intuitive to the user.

6.1. Topic word clouds

Word clouds are a graphical representation of a vocabulary where word size is based on its significance according to some metric. This is a commonly used visualisation technique for text where we want to emphasise some terms over others. In topic modelling, this value can be the word probability, lift, relevance or other metric. This can help give a quick impression of many of the most important words relating to a topic as measured by one of the metrics we have described.

Figure 3 shows word clouds of some topics from the LDA model trained on the Finnish NLF dataset.

6.2. LDAVis visualisation

LDAVis [Sievert and Shirley, 2014] is an interactive visualisation of LDA models that shows a plot of topics in relation to each other in a 2D space and the prevalence of the topic in the corpus. The plot is based on the PCA decomposition of the topic-term distributions. It also shows the top words of a selected topic with a slider that allows the user to adjust the λ parameter of the relevance metric, resulting in pure term-word probability weighting, pure lift weighting or anything in between. LDAVis is useful for exploring topics because it gives the user the freedom to try out different topic representations interactively.

The interactive component is implemented using Javascript and the entire visualisation is given to the user as an HTML file. In Figure 4 shows the LDAVis visualisation of one the Finnish NLF topics about the Finnish legislative assembly.

As the name implies, this visualisation is designed for LDA-type topic models and not easily adaptable for other topic models where we want to display topics that are aligned on some aspect such as language or time such as DTM.

6.3. Dynamic topic model visualisation

Even more than LDA, DTM greatly benefits from visualisation because it captures aspects of the corpus that are important to convey to the user in an intuitive manner.

There are currently not many established methods of visualising the outputs of dynamic topic modelling.

Figure 3: Word clouds using word probability of some topics from the Finnish NLF LDA model. Clockwise from upper left: Topics on the (1) Finnish legislative assembly, (2) house fires, (3) church and clergy and lastly, (4) the language debate.

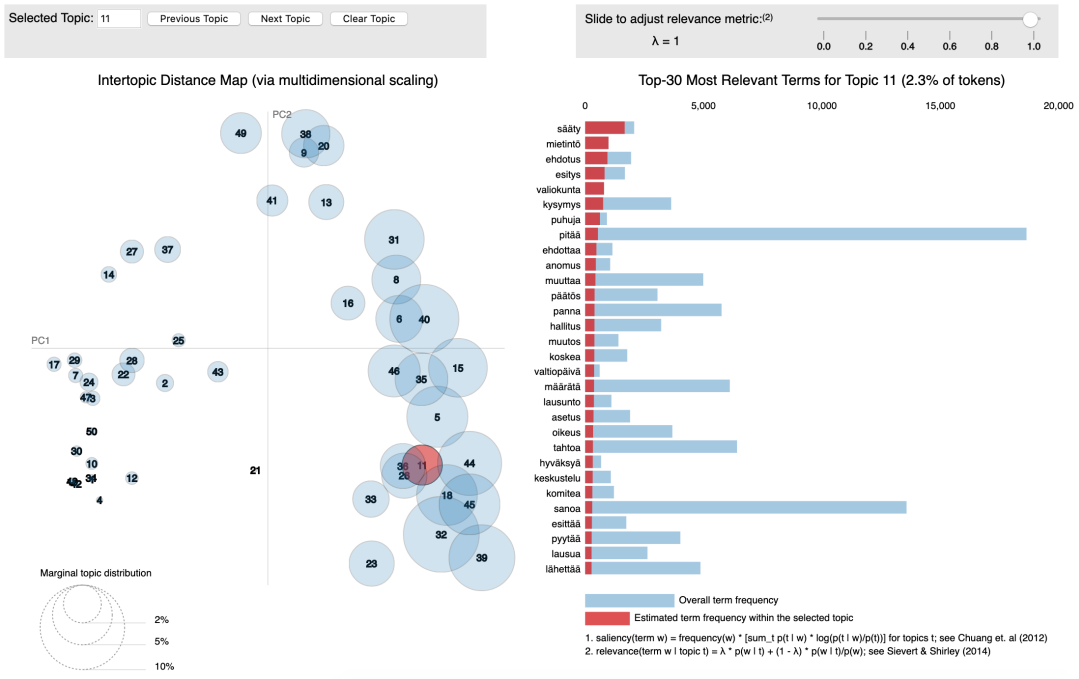


Figure 4: Screenshot of the LDAVis representation of a topic on the Finnish legislative assembly.

Papers typically show top topic words per time slice (as in Tables 9 and 8) or show the change in prominence of certain words over time (as in Figure 8). We present two other visualisation methods: heatmaps to show topic evolution, in Section 6.3.2, and stacked bar charts to show change in topic prominence, in Section 6.3.1). These are less commonly used (never, to our knowledge, reported in peer-reviewed papers), but have been helpful in providing us insights into the results of the topic modelling.

6.3.1. Topic prominence

We can compute the prominence of a topic in a time slice by adding up the topic proportions of all the documents in the time slice for that topic. We do this for all topics and normalise the result such that we can compare topic prominences across time. By plotting this in a bar chart or line plot, we can see how a topic's prominence rises and falls over time.

Figure 5 shows on a bar chart the relative prominence of all topics of the DTM trained on the NLF Finnish data as they change over the period covered. Since topic distribution changes over time, the top words of a topic also change, though this is not visible in the figure. The plot shown here is interactive – the user can hover their mouse on a time slice and see the top words for that topic in that time slice.

Figures 6 and 7 show as line plots the topic prominence for one topic and all topics in the trained DTM, respectively. A trade-off is apparent from these figures. While the first figure shows clearly the way the topic prominence changes, we cannot compare it to the other topics in the model. While the second figure gives us that information, it is hard to interpret when a large number of topics are shown together. In the NewsEye Demonstrator, we display the interactive bar chart (Figure 5) since it does a better job of conveying many of the details we get from a trained DTM in an easily intelligible manner.

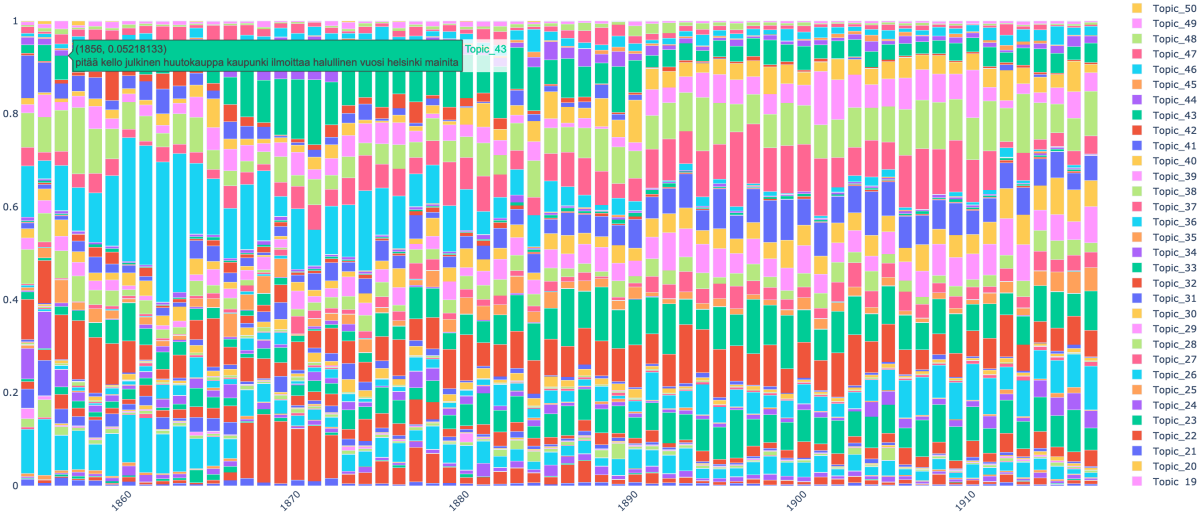


Figure 5: Bar chart of a DTM model trained on 64 years of Finnish newspapers showing topic prominence of 50 topics.

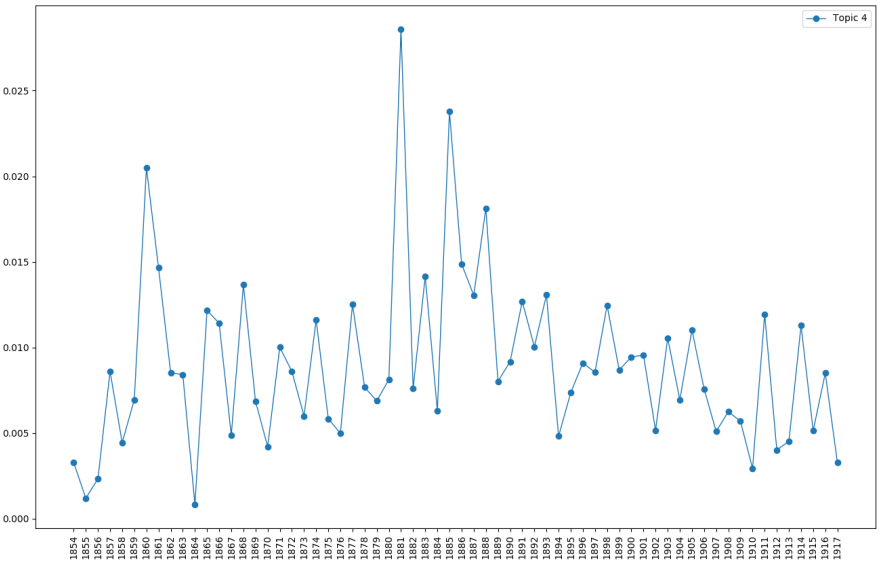


Figure 6: Topic prominence of Topic 4 from the Finnish news DTM.

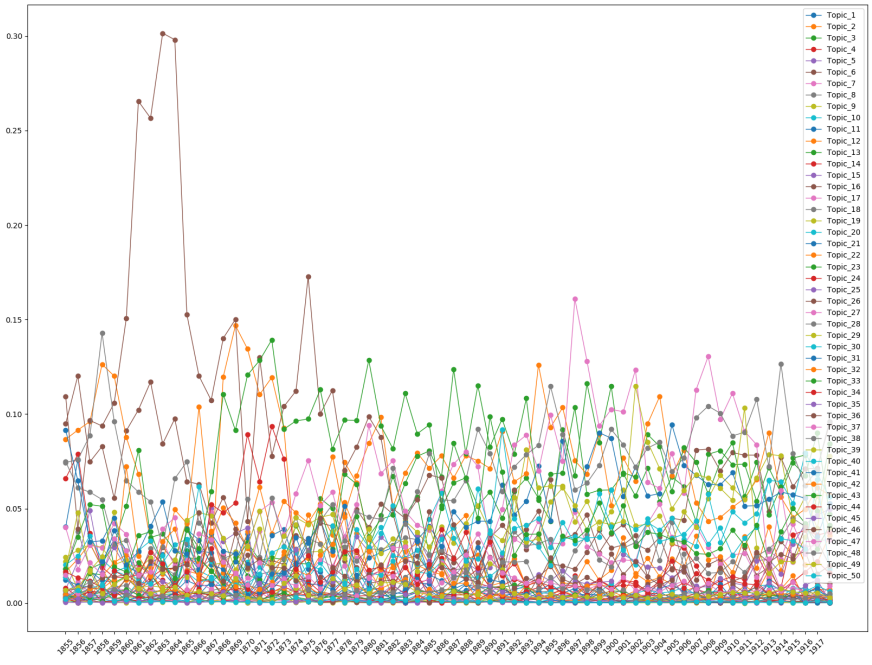


Figure 7: Topic prominence of all topics in the Finnish news DTM.

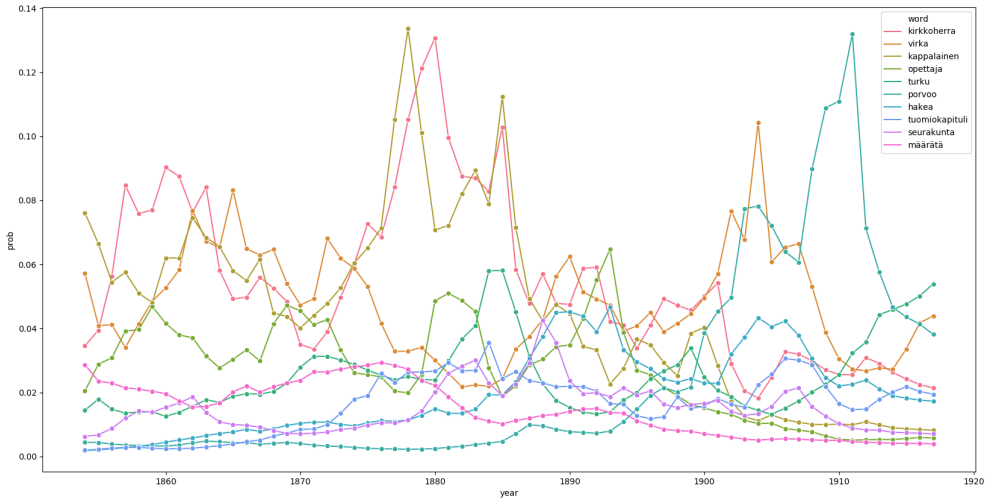


Figure 8: Plot of probability for some top words for a topic on the church and clergy in Finland

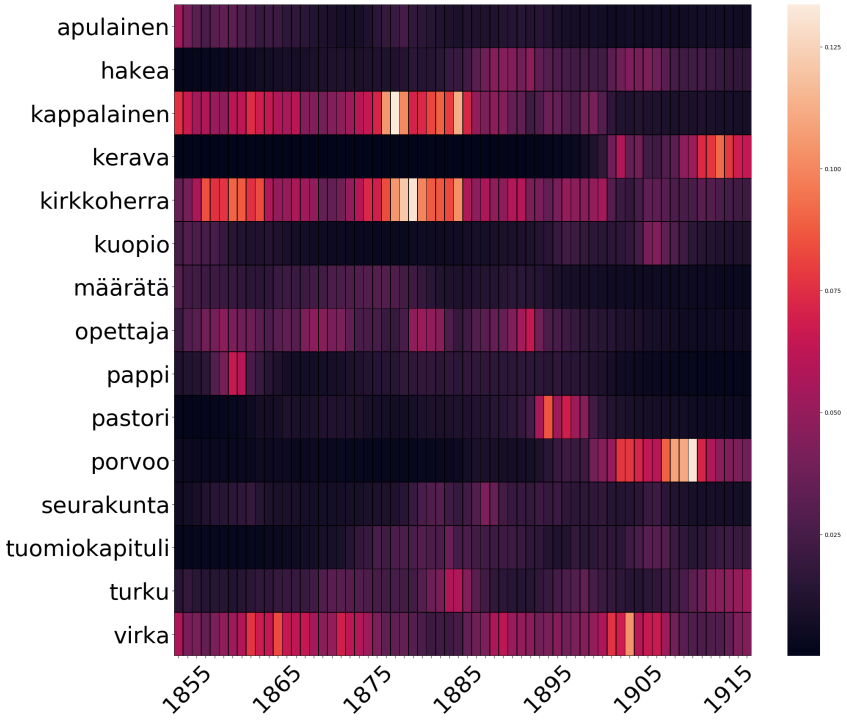


Figure 9: Heatmap showing the change in probability of some top words for a topic on the church and clergy in Finland. An alternative visualisation of the same information visualised in Figure 8

.

6.3.2. Topic evolution

If we want to focus on the topic evolution of a single topic and see more detail, we might be interested in how the significance of a word changes over time. Similar to topic prominence, we can also plot this significance over time. Figure 8 shows the probability of some prominent words for a topic about the church and clergy in Finland for 64 time slices in the form of a line plot. In this figure, how individual words rise and fall in prominence within the topic over time.

We also experimented with using heatmaps to illustrate topic evolution. Heat maps give a visual representation of matrices through colours, in our case the rows of the matrix are the top words of a topic and the columns are the significance of the words in each time slice. Figure 9 shows a heatmap of the same topic shown in Figure 8.

7. REST API

Some of the textual topic descriptions and visualisations discussed above are available for use by the NewsEye Demonstrator through the WP4 REST API[11]. Each item consists of a URI (e.g. `/lda/top-words`) and a specification of the parameters that the call accepts, possible HTTP response status codes and a description of the response body returned. The Demonstrator, or any other caller, can make an HTTP call to this URL, here always using the POST method, with the given parameters in order to perform the described analysis or lookup.

[11] https://newseye-wp4.cs.helsinki.fi

/lda/top-words

POST: Returns the top most probable words of a given topic.

This call is used for topics from an **LDA model**. A similar `top-words` call is provided for a **DTM model**

- `/dtm/top-words`

Body parameters:
`model_name` *string*: Name of trained model to describe
`topic_id` *string*: topic ID starting with 1
`time_slice` *int*: time slice starting with 1 or year (see `/dtm/valid-years`)
`lang` *string*: language code

Returned status codes:
`200`: The results are included in the response.
`404`: The specified topic does not exist.

Returned body (*application/json*):
`top_words` *string*: Top words of a topic according to their probability

/lda/topic-description

POST: Long description of given topic in a trained LDA model using extractive multi-document summarisation. At the moment we use the First Sentence summarisation method described in Section 5.2.1.

This call returns the topic description of a given topic in a **LDA model**.

Body parameters:
`model_name` *string*: Name of trained model to describe
`topic_id` *string*: Topic ID starting with 1

Returned status codes:
`200`: The results are included in the response.
`404`: The specified topic does not exist.

Returned body (*application/json*):
`topic_desc` *string*: Human-readable description of the topic

/lda/top-bigrams

POST: Top bigrams of given topic in a trained LDA model ranked by their topic relevance

This call returns the top topic bigrams of a given topic in an **LDA model**.

Body parameters:
`model_name` *string*: Name of trained model to describe
`topic_id` *string*: Topic ID starting with 1

Returned status codes:
`200`: The results are included in the response.
`404`: The specified topic does not exist.

Returned body (*application/json*):
`top_bigrams` *string*: Human-readable description of the topic in the form of bigrams

`/lda/word-cloud`

POST: Word cloud from a trained LDA model

This call returns a word cloud of a given topic in an **LDA model**. A similar `word-cloud` call is provided for a **DTM model**.

- `/dtm/word-cloud`

Body parameters:
`model_name` *string*: Name of trained model
`topic_id` *string*: topic ID starting with 1
`time_slice` *int*: time slice starting with 1 or year (see `/dtm/valid-years`)
`lang` *string*: language code

Returned status codes:
`200`: The results are included in the response.
`404`: The specified topic does not exist.

Returned body (*application/json*):
`topic_cloud` *image*: Word cloud of the specified topic in the specified language

`/lda/pyldavis`

POST: PyLDAVis visualisation of a trained LDA model

This call returns an HTML file containing the PyLDAVis visualisation of the trained LDA model.

Body parameters:
`model_name` *string*: Name of a trained LDA model to visualise

Returned status codes:
`200`: The results are included in the response.
`404`: The specified model does not exist or is not an LDA model

Returned body (*application/json*):
`pyldavis` *html*: The HTML file output of the PyLDAVis visualisation library.

`/dtm/bar-chart`

POST: Interactive bar chart visualisation of a trained DTM model

This call returns an HTML file containing the visualisation of the trained DTM model.

Body parameters:
`model_name` *string*: Name of a trained DTM model to visualise

Returned status codes:
`200`: The results are included in the response.
`404`: The specified model does not exist or is not an LDA model

Returned body (*application/json*):
`bar_chart` *html*: The HTML file output with the bar chart and interactive features.

8. Use by Digital Humanities collaborators

We worked with the University of Helsinki DH group (UH-DH) on a paper about exploring discourse dynamics in nineteenth-century Finnish newspapers. Our work examines discourses and discussions that were popular in the past but have since disappeared due to a variety of factors. We found that the heatmaps and barplots were especially useful in visualising a dynamic topic in an intuitive manner. This work was presented last year at the Digital Humanities in the Nordic Countries Conference (DHN 2020) and the manuscript has been accepted for publication in the post-conference proceedings (manuscript is attached to this deliverable).

9. Stance evolution

Stance detection task analyses writers' opinions towards given named entities. As an extension of stance detection task, stance evolution targets to visualise the change of the stance over various parameters such as time, media and country. The visualisation tool will provide data usable directly by end-users through the NewsEye demonstrator.

Given starting and ending years, the tool returns a list of NEs whole stances frequently or suddenly change between the starting and ending years of the whole corpus. This list of NEs is selected relying on a variance of yearly-stance polarities, which allows to measure how far a set of stance polarities is spread out from their average value.

Particularly, given a NE (i), for each year between the starting year (y_0) and ending year (y_n), we calculate the number of stances towards this NE of the whole corpus. Its yearly-stance polarity $(pol(i))$ is computed as shown in Equation (4) by dividing the difference in the number of positive stances $(pos(i))$ and negative stances $(neg(i))$ to the maximum number of stances.

$$pol(i) = \frac{pos(i) - neg(i)}{max_{j=y_0}^{y_n} pos(j) + neg(j) + neu(j)} \tag{4}$$

where $neu(i)$ as the number of neutral stances towards the NE (i).

Figure 10: Yearly-stance polarities over years, more nearly 1.0 is more positive.

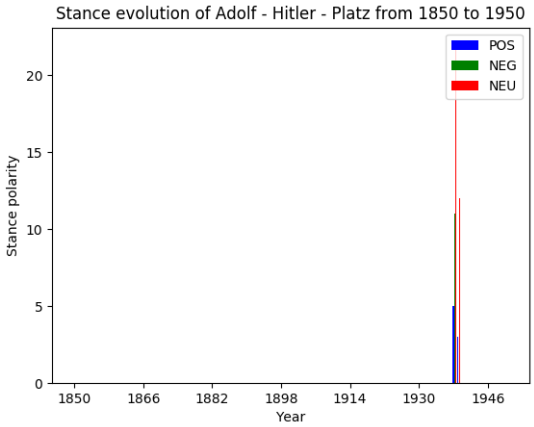


Figure 11: Statistics of positive / negative / neutral stances for each year.

Similarly, we compute yearly-stance polarities of other NEs and the variances of yearly-stance polarities. We keep the list of NEs whose variances are not 0s and ignore the others.

From the list, users can choose their interesting NE. Another option is that users can select any NE existing in the corpus. The tool, then, show two charts (line chart and bar chart) of stances towards the selected named entity between the starting and ending years. While the line chart indicates the yearly-stance polarities over years, more nearly 1.0 is more positive, the bar chart illustrates the number of positive / negative / neutral stances for each year. Examples of two graphs describing stance evolution towards the NE (Adolf - Hitler - Platz) between 1850 and 1950 are shown in Figures 10 and 11.

This tool potentially extends to visualise the stance towards other targets, pre-defined by DH scholars or other users. Some terms and concepts are known to be polarising and particularly help to answer their opinion-related research questions.

10. Conclusion

Textual topic labels. We experimented with different methods for generating textual topic labels. We demonstrate a method that scores words and phrases using different metrics and using the top scoring

words and phrases as a topic label. We show how sequence-to-sequence (seq2seq) models can be used to directly generate labels without the need for ranking candidate labels. Finally, we propose an ontological mapping method that produces topic labels in multiple languages and works in a zero-shot setting–it can be applied to languages that are not seen during training. This is especially useful when working with multilingual document collections such as in NewsEye.

Textual topic summaries. We experimented with different extractive multi-document summarisation methods to get more detailed topic descriptions that might provide users with more insight and context for understanding a topic.

Dynamic topic visualisation. We showed different visualisation methods for dynamic topic models and dynamic topics that shows the change of topic prominence and topic evolution over time.

LDA visualisation. We used the LDAVis library to generate interactive visualisations of trained LDA models that allows the user to adjust parameters to change the top topic words.

API. Some of the methods discussed here are now available for use by other work packages through the WP4 REST API.

Stance evolution. The work on stance evolution aims to visualise the evolution of attitude towards a named entity over a period of time.

Collaboration. We are collaborating with DH scholars in the NewsEye consortium on a paper that used some of the visualisations shown in this deliverable.

References

David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

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

Carson Sievert and Kenneth Shirley. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70. Association for Computational Linguistics, 2014.

Matt Taddy. On estimation and selection for topic models. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 1184–1193. PMLR, 2012.

Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 490–499, 2007.

Areej Alokaili, Nikolaos Aletras, and Mark Stevenson. Automatic generation of topic labels. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1965–1968, 2020.

Cristian Popa and Traian Rebedea. BART-TL: Weakly-supervised topic label generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1418–1425, 2021.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

Kaushal Kumar Maurya, Maunendra Sankar Desarkar, Yoshinobu Kano, and Kumari Deepshikha. ZmBART: An unsupervised cross-lingual transfer framework for language generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2804–2818, Online, Au-

gust 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.248. URL https://aclanthology.org/2021.findings-acl.248.

STT. Finnish news agency archive 1992-2018, source (http://urn.fi/urn:nbn:fi:lb-2019041501), 2019.

STT, Helsingin yliopisto, and Khalid Alnajjar. Finnish News Agency Archive 1992-2018, CoNLL-U, source (http://urn.fi/urn:nbn:fi:lb-2020031201), 2020. URL http://urn.fi/urn:nbn:fi:lb-2020031201.

Jenna Kanerva, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. Turku neural parser pipeline: An end-to-end system for the conll 2018 shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics, 2018.

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. Automatic labelling of topics with neural embeddings. *arXiv preprint arXiv:1612.05340*, 2016.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*, 2019.

Amparo Elizabeth Cano Basave, Yulan He, and Ruifeng Xu. Automatic labelling of topic models learned from Twitter by summarisation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 618–624, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-2101. URL https://www.aclweb.org/anthology/P14-2101.

A. Nenkova and Lucy Vanderwende. The impact of frequency on summarization. 2005.

Rada Mihalcea and Paul Tarau. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-3252.

Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1):107–117, 1998. ISSN 0169-7552. doi: https://doi.org/10.1016/S0169-7552(98)00110-X. URL https://www.sciencedirect.com/science/article/pii/S016975529800110X. Proceedings of the Seventh International World Wide Web Conference.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://www.aclweb.org/anthology/P02-1040.

A. Manuscript: Topic Modelling Discourse Dynamics in Historical Newspapers

Topic Modelling Discourse Dynamics in Historical Newspapers

Jani Marjanen[1*][0000−0002−3085−4862], Elaine Zosa[2*][0000−0003−2482−0663], Simon Hengchen[3][0000−0002−8453−7221], Lidia Pivovarova[2][0000−0002−0026−9902], and Mikko Tolonen[1][0000−0003−2892−8911]

[1] Helsinki Computational History Group, University of Helsinki
[2] Department of Computer Science, University of Helsinki
[3] Språkbanken, University of Gothenburg[‡]
firstname.lastname@{helsinki.fi,gu.se}

Abstract. This paper addresses methodological issues in diachronic data analysis for historical research. We apply two families of topic models (LDA and DTM) on a relatively large set of historical newspapers, with the aim of capturing and understanding discourse dynamics. Our case study focuses on newspapers and periodicals published in Finland between 1854 and 1917, but our method can easily be transposed to any diachronic data. Our main contributions are a) a combined sampling, training and inference procedure for applying topic models to huge and imbalanced diachronic text collections; b) a discussion on the differences between two topic models for this type of data; c) quantifying topic prominence for a period and thus a generalization of document-wise topic assignment to a discourse level; and d) a discussion of the role of humanistic interpretation with regard to analysing discourse dynamics through topic models.

Keywords: discourse dynamics, Finland, historical newspapers, nineteenth century, topic modeling, topic modelling

1 Introduction

This paper reports our experience on studying discursive change in Finnish newspapers from the second half of the nineteenth century. We are interested in grasping broad societal topics, discourses that cannot be reduced to mere words, isolated events or particular people. Our long-lasting goal is to investigate a global change in the presence of such topics and especially finding discourses that have disappeared or declined and thus could easily slip away in modern research. We believe that these research questions are better approached in a data-driven way without deciding what we are looking for beforehand, though the choice of the most suitable techniques for such research is still an open problem.

In this paper we focus on developing methodology. Choosing available algorithms for analysis guides possible outcomes as they are designed to be operationalised in

[‡] SH was affiliated with the University of Helsinki for most of this work.
[*] Equal contribution.

certain ways. Approaching our goal with mere word counts is counterproductive due to the sparseness of the language and the variety of discourse realisations in a given text. Further, word counts are unreliable with historical data due to never ending language change, spelling variations and text recognition errors.

Thus, as many other papers in the area of digital humanities, we utilize topic modelling as a proxy to discourses. In particular, we apply the "standard" Latent Dirichlet Allocation model [3, LDA] and its extension the Dynamic Topic Model [2, DTM], which is developed specifically to tackle temporal dynamics in data. However, any model has its limitations and tends to exaggerate certain phenomena while missing other ones. We focus on the difference between models and try to reveal their limitations in historical data analysis from the point of view that is relevant for historical scholarship.

Our main contributions are the following:

- We propose a **combined sampling, training and inference procedure** for applying topic models to large and imbalanced diachronic text collections.
- We discuss differences between two topic models, paying special attention to how they **can be used to trace discourse dynamics**.
- We propose a method to quantify **topic prominence for a period** and thus to generalize document-wise topic assignment to a discourse level.
- We **acknowledge and discuss the drawbacks of topic stretching,** which is typical for DTM. It is commonly known that DTM sometimes represents topics beyond the time period, but thus far there is no discussion in how researchers should tackle this for humanities questions.

In order to illustrate the appropriateness of the proposed methodology we discuss two use cases, one relating to discourses on church and religion and one that relates to education. The role of religion and education has been studied extensively in historical scholarship but there are no studies that deal with these topics through text mining of large-scale historical data. These two topics were chosen due to the the fact that the former was in general a discourse in decline relating to the process of secularization in Finnish society, whereas the latter increased in the second half of the nineteenth century and relates to the modernization of Finnish society and the inclusion of a larger share of the population in the sphere of basic education. In addition to these two interlinked discursive trends, we also use other examples to illustrate the strengths and weaknesses of LDA and DTM for this type of historical research.

2 Data

Our dataset is from the digitised newspaper collection of the National Library of Finland (NLF). This dataset contains articles from *all* newspapers and most periodicals that have been published in Finland from 1771 to 1917. Several studies have used parts of this dataset to investigate such issues as the development of the public sphere in Finland, the evolution of ideological terms in nineteenth-century Finland and the changing vocabulary of Finnish newspapers [36, 17, 16, 11, 21, 22, 25, 29, 12].

The full collection includes articles in Finnish, Swedish, Russian, and German. In this work we focus only on the Finnish portion starting from 1854 because this is the point where we determined we have sufficient yearly data to train topic models. The resulting subset has over 3.6 million articles and is composed of over 2.2 billion tokens. Figure 1a shows that the number of tokens published per year in Finnish-language papers increased steadily. The average article has 526 tokens but article length varies widely from year to year, as seen in Figures 1b and 1c which show the average article length and the number of articles per year. As made clear by these figures, there is a noticeable difference in the number of articles and average article length after 1910. This shift does not reflect the actual articles in the newspapers, but is the result of a change of OCR engine used to digitise the collection [20]. While the raw data is publicly available, we used the lemmatised version of the newspaper archive produced by Eetu Mäkelä, whom we thank.

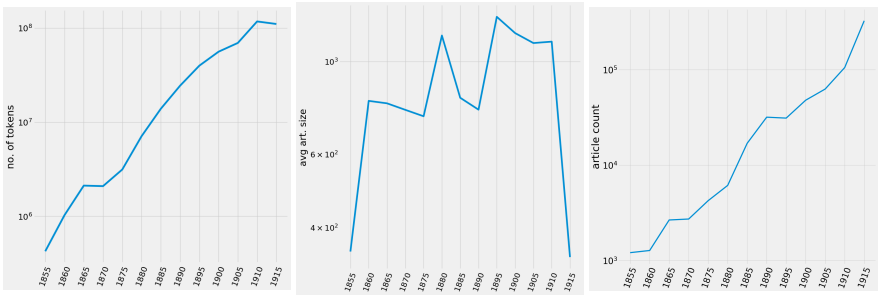
Still, even if the article segmentation differs in the latter period, Fig. 1a shows that there is steady increase in the vocabulary used in the Finnish-language newspapers published in the second half of the nineteenth century. They also covered more themes and regions. This entailed a process of diversification and modernization of the Finnish press, which has been widely discussed in historiography. As a collection, the newspapers vary a lot in style and focus. Some larger newspapers mainly contain political content, whereas others are rather specialised, and yet others thrived by giving a voice to the local public [35, 22, 16, 32]. This means that any analysis done on the entirety of the newspapers, like topic models, tend to balance out some of the differences between newspapers. This variety in the content, is also something that make newspapers such an interesting source material for historical research that is interesting in an overview of society. Although some issues were obviously not discussed because of taboo, courtesy or censorship, most of the themes present in public discourse are recorded in the newspapers and thus accessible to us in the present. Hence, we believe newspapers are an especially good source of assessing how the role of particular discourses changed over time.

2.1 Preprocessing the data

Given the size of the data and its inherent nature, notoriously the OCR quality and the unbalanced data from different time slices, we performed a series of pre-processing steps on the data.[1]

Despite prior work (albeit on English), showing that stemming has no real advantage for likelihood and topic coherence and can actually degrade topic stability [30], we follow [40, 10, 13] and use a lemmatised version of the corpus. Indeed, the work in [10] hints at the fact that Finnish, being much more inflected than English, would benefit from lemmatisation, whereas in [40, 13] the authors stem so as to reduce the huge number of token types due to OCR issues which impacts the performance of topic

[1]The more apt phrase "purposeful data modification", coined by [34], advocates that our material is not mere data that can go through a standardised "pre-processing" pipeline. Rather, the data is modified and altered only for the specific purposes of this study, and following this study's technical and scientific requirements only.



(a) Corpus size (b) Average article size (c) No. of articles

Fig. 1: Characteristics of the NLF dataset

modelling [38]. After lemmatisation, we remove tokens that occur less than 40 times in the collection, stopwords, punctuation marks and tokens with less than 3 characters. These are additional measures to further reduce the vocabulary size and mitigate the impact of OCR noise.

3 Topic Models

3.1 LDA

Topic modelling is an unsupervised method to extract topics from a collection of documents. Typically, a topic is a probability-weighted list of words that together express a theme or idea of what the topic is about. One of the most popular topic modelling methods currently in use is Latent Dirichlet Allocation (LDA), which is "a generative probabilistic model for collections of discrete data such as text corpora" [3]. It has been extensively used in the digital humanities to extract certain themes from a collection of texts [4]. In this model, a document is a mixture of topics and a topic is a probability distribution over a vocabulary. A limitation of LDA for historical research, in its vanilla form, is that it does not account for the temporal aspect of the data: every document in the collection is "considered synchronic", as time is simply not a variable in the model. Many document collections such as news archives, however, are diachronic—the documents are from different points in time, and scholars wish to study the evolution of topics.

There are different ways to overcome this limitation. One possibility is to split the data into time slices and train LDA separately on each slice. However, in this case LDA models for each slice would be independent of each other and there is no straightforward approach of matching topics from independent models trained on disjoint data. Another possibility, which we explore in this paper, is to train a single model for a subset of the whole data set over the entire time period and then use *topic prominence* as proxy for the dynamics of discourses over time.

To do this, we compute the prominence of a topic in a given year by summing up the topic contribution for each document in that year and then normalise this number by the sum of all topic contributions from all topics for that year, as in Equation 1.

$$P(z_k|y) = \frac{\sum_{j=1}^{|D_y|} P(z_k|d_j)}{\sum_{i=1}^{T} \sum_{j=1}^{|D_y|} P(z_i|d_j)} \quad (1)$$

where y is a year in the dataset, k is a topic index, D_y is the number of documents in year y, d_j is the j^{th} document in year y and T is the number of topics in the model.

The large size of the collection and its unbalanced nature is a problem for training topic models. It is computationally expensive to train a model with millions of articles and the resulting model would be heavily biased towards the latter years of newspaper collection because it has far more data. To overcome these issues, we sampled the collection such that we have a roughly similar data size for each year of the collection and as a result, we also get a vastly reduced dataset. However, to have a model of discourse dynamics that reflects the collection more closely, we compute topic prominence using the entire collection and not just the sampled portion. We do this by inferring the topic proportions of all the documents in the collection and using these inferred distributions to compute topic prominence.

3.2 DTM

As mentioned above, there are topic models that explicitly take into account the temporal dynamics of the data. One such model is the dynamic topic model (DTM). DTM is an extension of LDA that is designed to capture dynamic co-occurence patterns in diachronic data. In this model, the document collection is divided into discrete time slices and the model learns topics in each time slice with a contribution from the previous time slice. This results in topics that evolve slightly–words changing in saliency in relation to a topic–from one time step to the next.

However, DTM also has its own limitations. It is based on an assumption that each topic should be to some extent present in each time slice, which is not always the case with real-world data such as news archives where events and themes can sometimes disappear and then re-appear at some point in the future.

Perhaps more importantly for historical research, a weakness of DTM lies in its design: to accomplish alignment across time the topic model is fit across the whole vocabulary and thus smoothing between time slices is applied. As a result, events end up being "spread out" before and after they are known to happen. This problem only becomes evident after a thorough analysis: similar models in different fields such as lexical semantic change present the same issue – the dynamic topic model SCAN [7] generates a "plane" top word for the year 1700 (two centuries ahead of the Wright Flyer, and well before the word's first attested sense of "aeroplane"), while similar model GASC [26, 23] encounters the same weakness when modelling Ancient Greek. There is unfortunately no easy way to bypass this obstacle, which is particularly problematic when studying historical themes.

For both the LDA and DTM models, we use the Gensim implementation [28] with default model hyperparameters.

4 Related Work

Topic models are widely used in the digital humanities and social sciences to draw insights from large-scale collections [4] ranging from newspaper archives to academic journals. In this section, which we do not claim to be exhaustive, we discuss some of the previous works that aimed to capture historical trends in large data collections or used such collections to study discourses using topic models. All in all, these examples highlight that there is a need to discuss how topic models can be used to capture discursive change.

In [24] the authors use Latent Semantic Analysis, another topic modelling method, to study historical trends in eighteenth-century colonial America with articles from the *Pennsylvania Gazette*. Their work also used topic prominence to show, for instance, an increased interest in political issues as the country was heading towards revolution. The authors of [40] fit several topic models on Texan newspapers from 1829 to 2008. To discover interesting historical trends, the authors slice their data into four time bins, each corresponding to historically relevant periods. Such a slicing is also carried out in [9], where the author fits LDA models on Dutch-language Belgian socialist newspapers for three time slices that are historically relevant to the evolution of workers rights, with the aim of generating candidates for lexical semantic change.

Topic modelling has also been used in discourse analysis of newspaper data. In [37] the authors applied LDA to a selection of Italian ethnic newspapers published in the United States from 1898 to 1920 to examine the changing discourse around the Italian immigrant community, as told by the immigrants themselves, over time. They proposed a methodology combining topic modelling with close reading called discourse-driven topic modelling (DDTM). Another study examined anti-modern discourse in Europe from a collection of French-language newspapers [5]. In this case, however, the authors primarily use LDA as a tool to construct a sub-corpus of relevant articles that was then used for further analysis. Modernization was also an issue in the study of Indukaev [14], who uses LDA and word embeddings to study changing ideas of technology and modernization in Russian newspapers during the Medvedev and Putin presidencies.

LDA was not designed for capturing trends in diachronic data and so several methods have been developed to address this, such as DTM, Topics over Time [39, TOT], and the more recent Dynamic Embedded Topic Model [6, DETM], an extension of DTM that incorporates information from word embeddings during training. As far as we are aware, DTM and TOT have not been used for historical discourse analysis or applied to large-scale data collections. In the original papers presenting these methods, DTM was applied to 30,000 articles from the journal *Science* covering 120 years and TOT was applied to 208 State of the Union Presidential addresses covering more than 200 years. This was to demonstrate the evolution of scientific trends for the former and the localisation of significant historical events for the latter. Recently DETM was applied on a dataset of modern news articles about the COVID-19 pandemic where the authors observed differences between countries in how the pandemic and the reactions to it were framed [19].

In the mentioned cases researchers tackle the interpretative part of using topic models for humanistic research in different ways. Like Pääkkönen and Ylikoski [27] state, they toggle between some sort of topic realism, that is, using topic models to grasp

something that exists in the data, and topic instrumentalism, that is, using topic models to find something that can be further studied. Only Bunout [5] is a clear case of topic instrumentalism. All the other studies depart from some sort of realist position, and attempt to grasp policy shifts, ideas, discourses or framings of topics through topic models, but end up with correctives of some kind by highlighting the interpretative element [24, 37], by deploying formal evaluation by historians [9] or by using other quantitative methods to fine tune the results [14]. The interpretative aspect seems especially important when it comes to deciding on what researchers use the topics to study as they can reasonably relate to historical discourses, the semantics of related words, or simply ideas. How the topics are seen to represent these or, more likely, how the researchers use the topics to make an interpretation about these based on the topics, requires a strong element of interpretation [27]. Studies show that interpreters prefer to be able to go back to actual texts in order to make sense of topics [18], which is more than reasonable, but it also seems that there is a further need for researchers to understand how different topic-modelling methods represent diachronic data. Without this knowledge it is difficult to assess to which degree and for which time periods researchers need to manually assess individual documents.

5 Use Cases

What a discourse is, has been heavily theorised within the different strands of discourse analysis [1], but the advent of digital methods that can handle large textual data sets require quite some adjustment of discourse analysis as we know it. Like this article, others have turned to topic models to grasp changes in discourse [37, 5], but this article seeks specifically to discuss the interpretation that is required when we use topic models to study discourse dynamics. The probabilistic topic models set clear boundaries between topics and in doing so might merge or separate things that historians might regard as coherent topics. However, where the probabilistic model enforces boundaries, human interpretation in general is very bad at setting those boundaries and usually just identifies the core of a discourse or topic, but cannot say where it ends.

To get at the tension between topics and discourses, we approached the material without a predefined idea about which topics we wanted to study in order to keep the study as data-driven as possible. Our interest was to use topic modelling to capture topics that could in a meaningful way be related to societal discourses, that is themes that cannot be narrowed down to individual words, but still are reasonably coherent and form at least loose topics. To this end, we trained topic models with $k \in \{30; 50\}$, inferred topic distributions for the whole collection and inspected models by carefully going through the top words in each topic and using PyLDAVis[2] [31] to study overlap between topics and salience of terms per topic in LDA and heatmap visualizations for DTM. All topics were annotated and evaluated from the point of view of historical interpretation. We then opted to use the 50-topic model to study discourse changes over time. As is common, a portion of the topics seemed incoherent or were clearly the result of the layout in newspapers (e.g. boilerplate articles about prices etc.) and

[2] `https://github.com/bmabey/pyLDAvis`

did not produce interesting information about societal discourses. Further, some of the topics clearly overlap, so that a cluster of 2-5 topics can reasonably be seen as related to a particular societal discourse. The advantage of choosing 50 topics over 30 lies precisely in the possibility of merging topics later on in interpretation, while splitting them is more difficult.

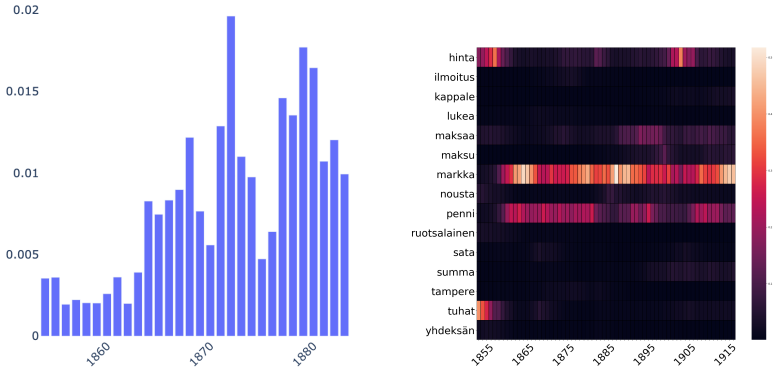
To discuss the benefits of LDA and DTM, we chose to focus on two specific themes, the discourse relating to religion and religious offices, and education. They are both rather neatly identifiable in the data, but display different trends. The former is in decline over the period of interest, whereas the latter increases in topic prominence. They can also be related to large scale processes in Finland, religious discourse to the secularization of society and education to the modernization of civic engagement.

5.1 DTM and Stretching of Topics

The two topic modelling methods perform in somewhat different ways. As mentioned, DTM is designed to incorporate temporal change in the topics, which means it includes a stronger sense of continuity in its representations of data. Whether or not this is desirable, depends on the research question, but our contention is that for studies interested in discursive change, this is either a problem or at least it is something that needs to be factored in making the historical interpretation. If we want to understand when certain discourses became dominant, declined, or even disappeared, this type of stretching cannot be allowed.

An exceptionally illustrative example of stretching among our fifty topics, is an introduction of the Finnish mark as a currency (Fig. 2a). With top words such as “mark”, “penny”, “price”, “thousand”, “pay” etc. the topic comes across as one with high internal coherence. We also see that the topic grows in prominence over time, from being relatively modest in the 1850s to gradually increased prominence after 1860. This makes sense, as the mark was adopted as currency in the year 1860 and after that self-evidently figured in public discourse. However, when we look at a heatmap visualization of the topic (Fig. 2b), we see how the topic stretches from the period 1854–1859 to the period 1860–1917, that is, from the period before the introduction of the mark to the period it was in use. After 1860 the words “mark” and “penny” are by far the most dominant terms in the topic, but for the period before 1860, the dominant terms are “price” and “thousand.” It is clear that “mark”, “penny”, “price”, and “thousand” are words that can belong to the same topic, but the heatmap representation clearly shows that the focus in the topic shifts. It is almost as if two related topics are merged as to represent one topic over the whole time period. In a situation where a historical interpretation highlights a change in past discourse, DTM produces continuity.

While there is obviously no right answer as to when one topic is stretched a bit or when different topics are simply merged together to provide a temporally continuous topic, it seems that DTM is especially problematic if one wants to study discourses that emerge or disappear in the middle of a time period studied. This means that any historical analysis using DTM requires a component of historical interpretation of not only topic coherence, but also topic coherence *over time*. Here, relying on word embeddings like in [14] can help, but this is primarily a task for evaluating the topics.



(a) Introduction of the Finnish mark in 1860 (y-axis indicates the topic probability)

(b) Heat map of terms linked to the introduction of the Finnish mark in 1860.

Fig. 2: Topic related to the introduction of the Finnish mark in 1860 (DTM). The most prominent terms in the heatmap are are "Mark" = *markka*, "penny" = *penni*, "price" = *hinta*, "thousand" = *tuhat*, "pay" = *maksu* and *maksaa*.

The speed of topic evolution can be controlled by a parameter in the DTM model. However, the 'ideal' amount of stretching is difficult to assess. For analysing discourse, this might in some cases be productive as it can point at links between nearby discourses, but is largely problematic as it hides discontinuities in the data. It becomes even problematic when dealing with material factors, like the introduction of the Finnish mark, as the stretching effect is likely to produce anachronistic representations, that is, placing something in the wrong period of time. Dealing with anachronism can perhaps be seen as one of the cornerstones of the historian's profession, which makes DTM as an anachronism prone method a poor match for historical study. Avoiding anachronisms completely is impossible, most historians would agree, but knowing when to avoid them and how to communicate about anachronistic elements in historical interpretation is key to history as a discipline [33].

5.2 Religion and Secularization

Our model performed well in grasping topics that relate to religion. The initial expectation regarding the discourse dynamics was that religious topics would be in decline. We hoped that using a topic model would be a way of showing this quantitatively. Results obtained from both LDA and DTM, presented in Figures 3a and 3b respectively, harmonize with our initial hypothesis, but do so differently. The DTM and LDA outputs cannot be aligned in any other way than manual interpretation by domain experts. In doing this we simply regarded topics that included several words that denote religious practices or offices as religious. Thus, the definition of "religious" is is rather narrow, but it also seems to match the topics that emerged from our data.

In order to inspect the discourse dynamics of religious topics, we have combined several topics that related to religious themes in the LDA model, whereas in the latter, DTM model, we only chose one topic to be represented.[3]

To our knowledge, topic models have not been used to study discursive change regarding secularization. However, in line with some earlier qualitative assessments [15], we hypothesize that this decline in religious discourse entails two interrelated developments: 1) Religion did not disappear from public discourse, but instead changed and disappeared from certain *types* of discourses. In the early nineteenth century, religion had a much more holistic presence in public discourse, meaning that religious metaphors and religious expressions and topics were used at a much vaster scale. 2) Over the course of the nineteenth century, religious topics became more focused. This means a segmentation of public discourse so that religious topics were increasingly confined to particular journals or genres.

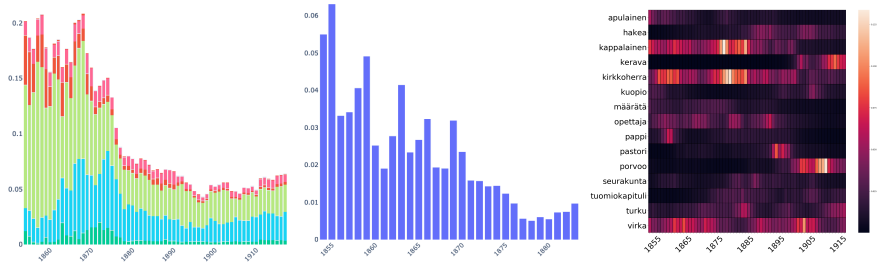
Keeping in mind the issue of stretching with DTM, we can look into the shifting saliency of words within the topic of religious offices and notice a shifting focus over time (Fig. 3c). In the early 1900s terms relating to "holding an office" and names of particular congregations become more dominant in the topic. This, again, suggests that DTM as a method does some stretching. There is a downside and an upside to this. On the one hand, the stretching distorts the topic prominence a bit by making it look like there is more continuity than in the LDA visualization. However, this may not be that crucial as the declining trends in Fig. 3a and Fig. 3b are rather similar. On the other hand, the stretching may be good for detecting conceptual links between different groups of words. In this particular case the stronger link between religious offices and some towns like Kerava and Porvoo, is probably indicative of a move of religious discourse from an overarching question to something that is more likely dealt with in conjunction to matters at local parishes. That is, religious offices were more often than before dealt with in connection to local congregations. This is in line with our above-mentioned assumption about religious discourse becoming more distinct.

5.3 Education and Modernity

While we expected religious themes to decline and become less central, we assumed there would be some themes that partly overlap with religion, but also would show an increasing trend. One example of this is the topic of education, which has historically been heavily interwoven with the church, but at the same time when basic education became available for a higher amount of people, it also became central in questioning the role of the church and religion. Education in nineteenth-century Finland was both central for ensuring conformity of the Lutheran faith, but paradoxically also was a vehicle of secularization. [8]

As in the case of religious discourse, alignment between DTM and LDA can only be made through human interpretation. It seems, that in this case DTM captures one topic

[3] We also experimented with more data-driven methods to cluster topics, including for example methods based on Jensen-Shannon Divergence. They unfortunately did not need to clusters that our domain experts would make sense of. Nonetheless, despite this, we still believe this is an interesting avenue to pursue which could help answer the common 'number of topics' question often brought up within the field.



(a) Topics related to religion on decline (LDA)

(b) Development of religious topic (chaplain, priest and office) over time

(c) Heatmap of terms linked to office of religion topic.

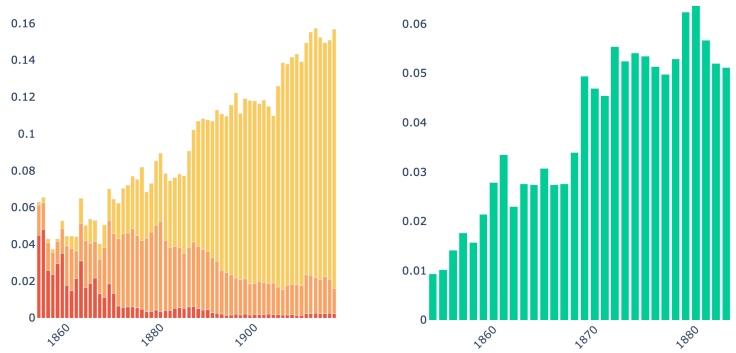
Fig. 3: Religious topics in LDA (a) and DTM (b,c); y-axis in (a, b) indicates the topics' probabilities. Most prominent terms in the heatmap are "chaplain" = *kappalainen*, "vicar" = *kirkkoherra*, "teacher" = *opettaja*, "priest" = *pappi*, "Porvoo" (a town), "parish" = *seurakunta*, "Turku" (a town), and "office" = *virka*.

that is fairly coherent, revolves around education and schooling, and is on the rise in the research period (Fig. 4b). For LDA, this is not the case, as an PyLDAVis inspection of most salient words across all fifty topics show that words like "school" and "folk school" appear mostly in three topics of which two are in decline and one heavily on the rise (Fig. 4a).

Interestingly, LDA and DTM seem to be pointing at a similar historical development. The two declining LDA topics are based on their most salient terms and are more focused on schools as buildings and institutions as well as teaching as a profession, whereas the topic on the rise includes salient vocabulary relating to, not only schools, but also meetings, civic engagements, and decision making. The DTM topic at hand shows a similar development which can be inspected in a heatmap of most salient terms over time. The terms "school", "child", and "teacher" dominate early in the period. By the end of the period the topic becomes broader, and terms like "municipality" and "meeting" have become more salient than the vocabulary relating to schools. Here the stretching of DTM creates the links that are also visible in the three LDA topics, and it shows a transformation in which educational issues are present in the whole topic, but focus shifts from concrete schools to civic engagement.

6 Conclusions

Our focus in this text has been on discourses that cannot be reduced to mere words, isolated events or particular people, but concern broader societal topics that either declined or gained in prominence. The interpretation of these topics and their contextualisation to nineteenth-century Finnish newspapers revealed clear topical cores that can be interpreted as an encouraging point of departure for further explorations based on topic models when aiming to understand Finnish public discourse through historical newspapers.



(a) Development of education topic over time (LDA)

(b) Development of education topic over time (DTM)

Fig. 4: Education topic in LDA and DTM; y-axis indicates the topics' probabilities

In this paper, we have learned that although it is difficult to pinpoint exactly where a discourse or topic ends, LDA and DTM can fairly reliably grasp many semi-coherent themes in past discourse and help us study the dynamics of discourses. However, our comparison of LDA and DTM as methods for getting at past discourse also shows that both methods require a very strong interpretative element in analysing historical discourses. DTM is much more prone to stretch or even merge topics, which requires an interpretative assessment of whether the stretching highlights interesting historical continuities or if it hides historical discontinuities that would require attention. We found that producing heatmaps of term saliency over time for each topic is a very useful way of doing this type of assessment. For LDA, stretching is not so much a problem, but often it seems interpretation is needed in seeing which topics logically relate to one another. While historical discourse analysis is traditionally tied strongly to a tradition of hermeneutic interpretation, the use of topic models to grasp discourse dynamics does not remove that need even if they allow for a quantification of discourse dynamics over time.

While we regard stretching in DTM as a predominantly negative feature, in some cases it can be useful. In the topics relating to education discussed above, the stretching in DTM actually points out links in discourses and is quite productive for the interpretative process of trying to figure out discourse dynamics. However, also in this case, the relevance of historical interpretation should be highlighted because it is very hard to tell whether the stretching of topics is an accurate reflection of the data or a shortcoming of the model. This can be addressed only by relating visualisations of topics to existing historical research and reading source texts. Humanities scholars are in general very good at making such interpretations, but it also needs to be noted that when we move further into the domain interpretative scholarship, we also lose some of the benefits of working with quantifying models. While it would be foolish to claim that a topic model represents data in a way that it provides simple facts about historical development, our use cases show that if we seek to find more reliable quantification LDA may

provide better results than DTM. Further, using LDA moves the interpretative stage further down in the research process, as it is likely to be about evaluating the connections between different topics over time. In DTM, the interpretation is likely moved forward to an evaluation of how well the algorithm did this merging topics. On this sense, our take on topic models harmonises with [27] who stress the role of humanistic interpretation, but for the sake of transparency suggest pushing the interpretation stage later in the research process.

Acknowledgements

This work has been supported by the European Union Horizon 2020 research and innovation programme under grants 770299 (NewsEye) and 825153 (EMBEDDIA). SH is funded by the project *Towards Computational Lexical Semantic Change Detection* supported by the Swedish Research Council (2019–2022; dnr 2018-01184).

References

1. Angermuller, J., Maingueneau, D., Wodak, R. (eds.): The discourse studies reader: Main currents in theory and analysis. John Benjamins Publishing, Amsterdam, the Netherlands ; Philadelphia PA (2014)
2. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the 23rd international conference on Machine Learning. pp. 113–120 (2006)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research **3**(Jan), 993–1022 (2003)
4. Brauer, R., Fridlund, M.: Historicizing topic models, a distant reading of topic modeling texts within historical studies. In: International Conference on Cultural Research in the context of Digital Humanities, St. Petersburg: Russian State Herzen University (2013)
5. Bunout, E.: Grasping the anti-modern discourse on Europe in the digitised press or can text mining help identify an ambiguous discourse? (2020)
6. Dieng, A.B., Ruiz, F.J., Blei, D.M.: The dynamic embedded topic model. arXiv preprint arXiv:1907.05545 (2019)
7. Frermann, L., Lapata, M.: A Bayesian model of diachronic meaning change. Transactions of the Association for Computational Linguistics **4**, 31–45 (2016)
8. Hanska, J., Vainio-Korhonen, K. (eds.): Huoneentaulun maailma: kasvatus ja koulutus Suomessa keskiajalta 1860-luvulle. Suomalaisen Kirjallisuuden Seuran toimituksia, 1266:1, Suomalaisen kirjallisuuden seura, Helsinki (2010), publication Title: Huoneentaulun maailma : kasvatus ja koulutus Suomessa keskiajalta 1860-luvulle
9. Hengchen, S.: When Does it Mean? Detecting Semantic Change in Historical Texts. Ph.D. thesis, Université libre de Bruxelles (2017)
10. Hengchen, S., Kanner, A.O., Marjanen, J.P., Mäkelä, E.: Comparing topic model stability between Finnish, Swedish, English and French. In: Digital Humanities in the Nordic Countries (2018)
11. Hengchen, S., Ros, R., Marjanen, J.: A data-driven approach to the changing vocabulary of the nation in English, Dutch, Swedish and Finnish newspapers, 1750-1950. In: Proceedings of the Digital Humanities (DH) conference (2019)
12. Hengchen, S., Ros, R., Marjanen, J., Tolonen, M.: A data-driven approach to studying changing vocabularies in historical newspaper collections. Digital Scholarship in the Humanities (2021)

13. Hill, M.J., Hengchen, S.: Quantifying the impact of dirty OCR on historical text analysis: Eighteenth century collections online as a case study. Digital Scholarship in the Humanities **34**(4), 825–843 (2019)
14. Indukaev, A.: Studying Ideational Change in Russian Politics with Topic Models and Word Embeddings. In: Gritsenko, D., Wijermars, M., Kopotev, M. (eds.) Palgrave Handbook of Digital Russia Studies. Palgrave Macmillan, Basingstoke (2021)
15. Juva, M.: Valtiokirkosta kansankirkoksi: Suomen kirkon vastaus kahdeksankymmentäluvun haasteeseen. WSOY, Porvoo (1960)
16. Kokko, H.: Suomenkielisen julkisuuden nousu 1850-luvulla ja sen yhteiskunnallinen merkitys. Historiallinen Aikakauskirja **117**(1), 5–21 (2019)
17. La Mela, M., Tamper, M., Kettunen, K.: Finding Nineteenth-century Berry Spots: Recognizing and Linking Place Names in a Historical Newspaper Berry-picking Corpus. In: Navarretta, C., Agirrezabal, M., Maegaard, B. (eds.) DHN 2019 - Digital Humanities in the Nordic Countries. pp. 295–307. CEUR Workshop Proceedings, CEUR (2019), https://cst.dk/DHN2019/DHN2019.html
18. Lee, T.Y., Smith, A., Seppi, K., Elmqvist, N., Boyd-Graber, J., Findlater, L.: The human touch: How non-expert users perceive, interpret, and fix topic models. International Journal of Human-Computer Studies **105**, 28–42 (Sep 2017). https://doi.org/10.1016/j.ijhcs.2017.03.007, https://linkinghub.elsevier.com/retrieve/pii/S1071581917300472
19. Li, Y., Nair, P., Wen, Z., Chafi, I., Okhmatovskaia, A., Powell, G., Shen, Y., Buckeridge, D.: Global surveillance of covid-19 by mining news media using a multi-source dynamic embedded topic model. In: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. pp. 1–14 (2020)
20. Mäkelä, E., Tolonen, M., Marjanen, J., Kanner, A., Vaara, V., Lahti, L.: Interdisciplinary collaboration in studying newspaper materiality. In: Krauwer , S., Fišer, D. (eds.) Twin Talks Workshop at DHN 2019. pp. 55–66. CEUR Workshop Proceedings, CEUR-WS.org, Germany (2019)
21. Marjanen, J., Pivovarova, L., Zosa, E., Kurunmäki, J.: Clustering ideological terms in historical newspaper data with diachronic word embeddings. In: 5th International Workshop on Computational History, HistoInformatics 2019. CEUR-WS (2019)
22. Marjanen, J., Vaara, V., Kanner, A., Roivainen, H., Mäkelä, E., Lahti, L., Tolonen, M.: A national public sphere? Analyzing the language, location, and form of newspapers in Finland, 1771–1917. Journal of European Periodical Studies **4**(1), 54–77 (2019)
23. McGillivray, B., Hengchen, S., Lähteenoja, V., Palma, M., Vatri, A.: A computational approach to lexical polysemy in Ancient Greek. Digital Scholarship in the Humanities **34**(4), 893–907 (2019)
24. Newman, D.J., Block, S.: Probabilistic topic decomposition of an eighteenth-century american newspaper. Journal of the American Society for Information Science and Technology **57**(6), 753–767 (2006)
25. Oiva, M., Nivala, A., Salmi, H., Latva, O., Jalava, M., Keck, J., Domínguez, L.M., Parker, J.: Spreading News in 1904: The Media Coverage of Nikolay Bobrikov's Shooting. Media History **26**(4), 391–407 (Oct 2020). https://doi.org/10.1080/13688804.2019.1652090, https://www.tandfonline.com/doi/full/10.1080/13688804.2019.1652090
26. Perrone, V., Palma, M., Hengchen, S., Vatri, A., Smith, J.Q., McGillivray, B.: GASC: Genre-aware semantic change for Ancient Greek. In: Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change. pp. 56–66. Association for Computational Linguistics, Florence, Italy (Aug 2019). https://doi.org/10.18653/v1/W19-4707, https://www.aclweb.org/anthology/W19-4707

27. Pääkkönen, J., Ylikoski, P.: Humanistic interpretation and machine learning. Synthese (Sep 2020). https://doi.org/10.1007/s11229-020-02806-w, http://link.springer.com/10.1007/s11229-020-02806-w
28. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010), http://is.muni.cz/publication/884893/en
29. Salmi, H., Paju, P., Rantala, H., Nivala, A., Vesanto, A., Ginter, F.: The reuse of texts in Finnish newspapers and journals, 1771–1920: A digital humanities perspective. Historical Methods: A Journal of Quantitative and Interdisciplinary History pp. 1–15 (Sep 2020). https://doi.org/10.1080/01615440.2020.1803166, https://www.tandfonline.com/doi/full/10.1080/01615440.2020.1803166
30. Schofield, A., Mimno, D.: Comparing apples to apple: The effects of stemmers on topic models. Transactions of the Association for Computational Linguistics **4**, 287–300 (2016)
31. Sievert, C., Shirley, K.: Ldavis: A method for visualizing and interpreting topics. In: Proceedings of the workshop on interactive language learning, visualization, and interfaces. pp. 63–70 (2014)
32. Sorvali, S.: "Pyydän nöyrimmästi sijaa seuraavalle" – Yleisönosaston synty, vakiintuminen ja merkitys autonomian ajan Suomen lehdistössä. Historiallinen Aikakauskirja **118**(3), 324–339 (2020)
33. Syrjämäki, S.: Sins of a historian: Perspectives on the problem of anachronism. Ph.D. thesis, Tampere University Press, Tampere (2011), oCLC: 816367378
34. Thompson, L., Mimno, D.: Authorless topic models: Biasing models away from known structure. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 3903–3914. Association for Computational Linguistics, Santa Fe, New Mexico, USA (Aug 2018), https://www.aclweb.org/anthology/C18-1329
35. Tommila, P., Landgrén, L.F., Leino-Kaukiainen, P.: Suomen lehdistön historia 1. Sanomalehdistön vaiheet vuoteen 1905. Kustannuskiila, Kuopio (1988)
36. Vesanto, A., Nivala, A., Rantala, H., Salakoski, T., Salmi, H., Ginter, F.: Applying BLAST to text reuse detection in finnish newspapers and journals, 1771-1910. In: Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language. pp. 54–58 (2017)
37. Viola, L., Verheul, J.: Mining ethnicity: Discourse-driven topic modelling of immigrant discourses in the USA, 1898–1920. Digital Scholarship in the Humanities (2019)
38. Walker, D., Lund, W.B., Ringger, E.: Evaluating models of latent document semantics in the presence of ocr errors. In: Proceedings of the 2010 conference on empirical methods in natural language processing. pp. 240–250 (2010)
39. Wang, X., McCallum, A.: Topics over time: a non-markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 424–433 (2006)
40. Yang, T.I., Torget, A., Mihalcea, R.: Topic modeling on historical newspapers. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. pp. 96–104 (2011)

B. Manuscript: Multilingual Topic Labelling of News Topics using Ontological Mapping

Multilingual Topic Labelling of News Topics using Ontological Mapping

Elaine Zosa[0000−1111−2222−3333], Lidia Pivovarova[1111−2222−3333−4444], Michele Boggia[2222−−3333−4444−5555], and Sardana Ivanova[2222−−3333−4444−5555]

University of Helsinki, Finland
firstname.lastname@helsinki.fi

Abstract. The large volume of news produced daily makes topic modelling useful for analysing topical trends. A topic is usually represented by a ranked list of words but this can be difficult and time-consuming for humans to interpret. Therefore, various methods have been proposed to generate labels that capture the semantic content of a topic. However, there has been no work so far on coming up with multilingual labels which can be useful for exploring multilingual news collections. We propose an ontological mapping method that maps topics to concepts in a language-agnostic news ontology. We show that our method performs on par with state-of-the-art label generation methods, is able to produce multilingual labels, and can be applied to topics from languages that have not been seen during training without any modifications.

Keywords: topic labelling · ontology linking · cross-lingual embeddings

1 Introduction

Topic models uncover the latent themes in a document collection through the co-occurrences of words in documents [4]. The large volume of news produced daily makes topic models especially useful for tracking and analysing news trends [12, 14, 17]. A topic is usually represented by a ranked list of words but these words might be difficult and time-consuming to interpret for humans [10]. Therefore various methods have been proposed to assign concise labels to topics to improve interpretability [1, 3, 16, 18]. However, there has been no work so far on coming up with multilingual topic labels. Generating labels in multiple languages allows users to compare topical trends across linguistic boundaries without having to align topics and to explore news collections by users who might not have the necessary linguistic skills to do otherwise.

In this work we are interested in assigning concise multilingual labels to news topics. We propose an ontological mapping method that maps topics to concepts in a language-agnostic news ontology. These concepts have labels in multiple languages that we use as topic labels. We approach ontology mapping as a multilabel classification task where a topic can be classified as belonging to multiple concepts.

We train our classifier on a dataset of Finnish news and test it on Finnish and English topics, using the distant supervision approach proposed in Ref. [1], where articles are used as training data. Our method produces results that are on par with state-of-the-art label generation methods, produces multilingual labels and can be used for topics in languages that have not been used during training without any modification. The contributions in this paper are: (1) an ontological mapping approach that can produce topic labels in multiple languages; (2) a method based on contextualised cross-lingual embeddings that works in a zero-shot setting, assigning labels to topics in languages not seen during training; and (3) a novel dataset of Finnish news topics with gold standard labels.[1]

2 Related Work

Several existing methods for automatic topic labelling generate candidate labels either by extracting short phrases from topic-related documents [2,9,16] or from external sources such as Wikipedia [1,9] and then ranking the candidates according to their relevance to the topic using distance metrics such as cosine distance [3] or the Kullback-Leibler divergence [8,16].

Wikipedia is a popular external corpora for topic labelling, using article titles as candidate labels [3,9]. However, Ref. [9] argues that the broad domain covered by Wikipedia might make it unsuitable for labelling topics from a domain-specific corpus, such as biomedical research papers. Moreover, Wikipedia sizes vary widely across different languages. Some previous work also used ontologies [5,7] but their methods rely on network analysis techniques to extract labels from the ontologies.

A more recent development is using deep learning to directly generate labels. Ref. [1] proposes a sequence-to-sequence model (seq2seq) trained on a synthetic dataset of Wikipedia articles and titles while Ref. [18] finetune BART, a pretrained transformer-based language model [11], with topic keywords and candidate labels from weak labellers to generate labels.

3 Experimental Setup

3.1 Models

Ontology Mapping. We propose an ontological mapping method that maps topics to concepts in a language-agnostic news ontology and use the corresponding labels for these concepts—available in multiple languages—as topic labels. We treat the ontology mapping problem as a multilabel classification task where a topic can be classified as belonging to one or more concepts in the ontology.

The classifier takes as an input a sequence $X = (x_1, \ldots, x_n)$ of the n top terms of a topic, and predicts $P(c_i|X)$, the probabilities for each ontology concept $c_i \in C$. The topic labels are obtained from the distribution $P(c_i|X)$ as follows: First, a list of label candidates is obtained by considering all c_i such that $P(c_i|X) > t$,

[1] Source code and dataset will be publicly available upon acceptance.

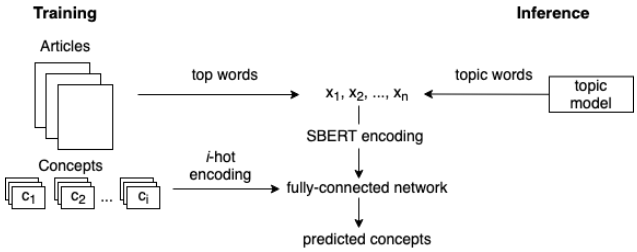


Fig. 1. News concepts prediction pipeline.

where t is the classification threshold. Then, we propagate the predicted concepts to the top of the ontology. For instance, if a topic is classified as belonging to concept 01005000:CINEMA, it also belongs to concept 01000000:ARTS, CULTURE AND ENTERTAINMENT, the parent of 01005000:CINEMA. Lastly, we obtain the topic labels by taking the most frequent concepts among the candidates and taking the labels of these concepts in the preferred language.

To compute the probabilities $P(c_i|X)$, we encode the top terms $(x_1, \ldots, x_n)$ using SBERT [19][2] and pass this representation to a classifier composed of two fully-connected layers with a ReLU non-linearity and a softmax activation. We set the classification threshold t to 0.03 as determined by the validation set. We refer to this as the **ontology** model. We illustrate this model on Figure 1.

Comparisons to State-of-the-art. We also investigate how our ontology mapping method compares to methods that directly generate topic labels. Ref. [1] uses an RNN-based encoder-decoder architecture with attention as a seq2seq model while Ref. [18] finetunes a pretrained BART model. Both methods have reported state-of-the-art results on English topics from multiple domains.

We implement a RNN seq2seq model using the same hyperparameters as [1]: 300-dim for the embedding layer and a hidden dimension of 200. We refer to this as the **rnn** model. We also implement a slightly modified model where we replace RNN with transformers, which has yielded state-of-the-art results in many NLP tasks. We use the hyperparameters from the original transformers model [22]: 6 layers for the encoder and decoder with 8 attention heads and a embedding dimension of 512. We refer to this as the **transformer** model.

Instead of BART which is trained only on English, we finetune a multilingual version, mBART [13], and set the source and target languages to Finnish. We finetuned mBART-25 from HuggingFace[3] for 5 epochs. We use the AdamW optimizer with weight decay set to 0.01. We refer to this as the **mbart** model [4]. For consistency, all the models except mbart are trained using Adam optimizer for 30 epochs with early stopping based on the validation loss.

[2] We use the multilingual model *distiluse-base-multilingual-cased*.
[3] https://huggingface.co/facebook/mbart-large-cc25
[4] While the mBART encoder is in a multilingual space, it cannot be used directly for cross-lingual language generation [15].

3.2 Datasets

News Ontology. We use the IPTC Subject Codes as our news ontology.[5] This is a language-agnostic ontology designed to organise news content. Labels for concepts are available in multiple languages—in this work we focus specifically on Finnish and English. This ontology has three levels with 17 high-level concepts, 166 mid-level concepts and 1221 fine-grained concepts. Mid-level concepts have exactly one parent and multiple children.

Training Data. We use news articles from 2017 of the Finnish News Agency dataset [20, 21] which have been tagged with IPTC concepts and lemmatized with the Turku neural parser [6]. Following the distant-supervision approach in [1], we construct a dataset where the top n words of an article are treated as input $X = (x_1, \ldots, x_n)$ and the tagged concepts are the target C; an article can be mapped to multiple concepts. Top words can either be the top 30 scoring words by tf-idf (**tfidf** dataset) or the first 30 unique content words in the article (**sent** dataset). All models are tained on both datasets. For each dataset, we have 385803 article-concept pairs which we split 80/10/10 into train, validation and test sets.

Test Data. For Finnish topics, we train an LDA model for 100 topics on the articles from 2018 of the Finnish news dataset and select 30 topics with high topic coherence for evaluation. We also check that the topics are diverse enough such that they cover a broad range of subjects.

To obtain gold standard labels for these topics, we recruited three fluent Finnish speakers to provide labels for each of the selected topics. For each topic, the annotators received the top 20 words and three articles closely associated with the topic. We provided the following instructions to the annotators:

Given the words associated with a topic, provide labels (in Finnish) for that topic. There are 30 topics in all. You can propose as many labels as you want, around 1 to 3 labels is a good number. We encourage concise labels (maybe 1-3 words) but the specificity of the labels is up to you. If you want to know more about a topic, we also provide some articles that are closely related to the topic. These articles are from 2018.

We reviewed the given labels to make sure the annotators understood the task and the labels are relevant to the topic. We use all unique labels as our gold standard, which resulted in seven labels for each topic on average. While previous studies on topic labelling mainly relied on having humans evaluate the labels outputted by their methods, we opted to have annotators *provide* labels instead because this will give us an insight into how someone would interpret a topic[6]. During inference, the input X are the top 30 words for each topic.

To test our model in a cross-lingual zero-shot setting, we use the English news topics and gold standard labels from the NETL dataset [3]. These gold labels were obtained by generating candidate labels from Wikipedia titles and asking humans to evaluate the labels on a scale of 0-3. This dataset has 59 news

[5] https://cv.iptc.org/newscodes/subjectcode/

[6] Volunteers are compensated for their efforts. We limited our test data to 30 topics due to budget constraints.

	Prec	Rec	F-score
Finnish news			
baseline: top 5 terms	*89.47*	*88.08*	*88.49*
ontology-tfidf	94.54	95.42	94.95
ontology-sent	95.18	**95.96**	95.54
mbart-tfidf	93.99	94.56	94.19
mbart-sent	94.02	95.04	94.51
rnn-tfidf	**96.15**	95.61	**95.75**
rnn-sent	95.1	94.63	94.71
transformer-tfidf	94.26	94.42	94.30
transformer-sent	95.45	94.73	94.98
English news			
baseline: top 5 terms	***98.17***	***96.58***	***97.32***
ontology-tfidf	97.00	95.25	96.04
ontlogy-sent	97.18	95.43	96.21

Table 1. Averaged BERTScores between labels generated by the models and the gold standard labels for Finnish and English news topics.

topics with 19 associated labels but we only take as gold labels those that have a mean rating of at least 2.0, giving us 330 topic-label pairs. We use default topic labels—top five terms of each topic—as the baselines.

4 Results and Discussion

We use BERTScore [23] to evaluate the labels generated by the models with regards to the gold standard labels. BERTScore finds optimal correspondences between gold standard tokens and generated tokens and from these correspondences, recall, precision, and F-score are computed.

We show the average BERTScores for the Finnish news topics at the top of Table 1. All models outperform the baseline by a large margin which shows that labels to ontology concepts are more aligned with human-preferred labels than the top topic words. The rnn-tfidf model obtained the best scores followed by ontology-sent. The transformer-sent and mbart-sent models also obtain comparable results. We do not see a significant difference in performance between training on the tfidf or sent datasets. In Table 2 (top), we show an example of the labels generated by the models and the gold standard labels. All models give sufficiently suitable labels, focusing on motor sports. However only the ontology-sent model was able to get Formula 1 as one of its labels.

We also demonstrate the ability of the ontology models to label topics in a language it has not seen during training by testing it on English news topics from the NETL dataset [3]. This dataset was also used in Ref. [1] for testing but our results are not comparable since they present the scores for topics from all domains while we only use the news topics. The results are shown at the bottom of Table 1. Although the ontology models do not outperform the baseline, they are still able to generate English labels that are very close to the gold labels considering that it has only been trained on Finnish. From the example in Table 2

	Finnish topic
Topic	räikkönen, bottas, ajaa *(to drive)*, hamilto, mercedes
Gold	formula, formulat, formula 1, f1, formula-auto, aika-ajot *(time trial)*, moottoriurheilu *(motor sport)*
rnn-tfidf	autourheilu *(auto sport)*, urheilutapahtumat *(sports event)*, mm-kisat *(world championship)*, urheilu *(sport)*, urheilijat *(athletes)*
transformer-sent	urheilutapahtumat *(sports event)*, mm-kisat *(world championship)*, urheilu *(sport)*, autourheilu *(auto sport)*, kansainväliset *(international)*
mbart-sent	autourheilu moottoriurheilu, urheilutapahtumat, mm-kisat , urheilijat pelaajat, urheilu
ontology-sent	ID: 15000000, fi: urheilu, en: sport; ID: 15039000, fi: autourheilu moottoriurheilu, en: motor racing; ID: 15073000, fi: urheilutapahtumat, en: sports event; ID: 15039001, fi: formula 1, en: formula one; ID: 15073026, fi: mm-kisat, en: world championship
	English topic
Topic	film, movie star, director, hollywood, actor, minute, direct, story, witch
Gold	fantasy film, film adaptation, quentin tarantino, a movie, martin scorsese, film director, film
ontology-sent	ID: 01005001, en: film festival, fi: elokuvajuhlat; ID: 04010003, en: cinema industry, fi: elokuvateollisuus; ID: 08000000, en: human interest, fi: human interest; ID: 01022000, en: culture (general), fi: kulttuuri yleistä; ID: 04010000, en: media, fi: mediatalous

Table 2. Generated labels for selected topics. Finnish labels are manually translated except for ontology-sent. For ontology-sent, we provide the concept ID and the corresponding Finnish and English labels.

(bottom), we also observe that the gold labels are overly specific, suggesting names of directors as labels when the topic is about the film industry in general. We believe this is due to the procedure used to obtain the gold labels, where the annotators were asked to *rate* labels rather than propose their own.

5 Conclusion

We propose a straightforward ontology mapping method for producing multilingual labels for news topics. We cast ontology mapping as a multilabel classification task, represent topics as contextualised cross-lingual embeddings with SBERT and classify them into concepts from a language-agnostic news ontology where concepts have labels in multiple languages. Our method performs on par with state-of-the-art topic label generation methods, produces multilingual labels, and works on multiple languages without additional training. We also show that labels of ontology concepts correlate highly with labels preferred by humans.

Acknowledgements

This work has been supported by the European Union's Horizon 2020 research and innovation programme under grant 770299 (NewsEye) and 825153 (EMBEDDIA).

References

1. Alokaili, A., Aletras, N., Stevenson, M.: Automatic generation of topic labels. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1965–1968 (2020)
2. Basave, A.E.C., He, Y., Xu, R.: Automatic labelling of topic models learned from twitter by summarisation. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 618–624 (2014)
3. Bhatia, S., Lau, J.H., Baldwin, T.: Automatic labelling of topics with neural embeddings. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 953–963 (2016)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
5. Hulpus, I., Hayes, C., Karnstedt, M., Greene, D.: Unsupervised graph-based topic labelling using dbpedia. In: Proceedings of the sixth ACM international conference on Web search and data mining. pp. 465–474 (2013)
6. Kanerva, J., Ginter, F., Miekka, N., Leino, A., Salakoski, T.: Turku neural parser pipeline: An end-to-end system for the conll 2018 shared task. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Association for Computational Linguistics (2018)
7. Kim, H.H., Rhee, H.Y.: An ontology-based labeling of influential topics using topic network analysis. Journal of Information Processing Systems **15**(5), 1096–1107 (2019)
8. Kou, W., Li, F., Baldwin, T.: Automatic labelling of topic models using word vectors and letter trigram vectors. In: AIRS. pp. 253–264. Springer (2015)
9. Lau, J.H., Grieser, K., Newman, D., Baldwin, T.: Automatic labelling of topic models. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. pp. 1536–1545 (2011)
10. Lau, J.H., Newman, D., Baldwin, T.: 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 (2014)
11. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7871–7880 (2020)
12. Li, Y., Nair, P., Wen, Z., Chafi, I., Okhmatovskaia, A., Powell, G., Shen, Y., Buckeridge, D.: Global surveillance of covid-19 by mining news media using a multi-source dynamic embedded topic model. In: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. pp. 1–14 (2020)
13. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics **8**, 726–742 (2020)
14. Marjanen, J., Zosa, E., Hengchen, S., Pivovarova, L., Tolonen, M.: Topic modelling discourse dynamics in historical newspapers. arXiv preprint arXiv:2011.10428 (2020)

15. Maurya, K.K., Desarkar, M.S., Kano, Y., Deepshikha, K.: ZmBART: An unsupervised cross-lingual transfer framework for language generation. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 2804–2818. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.findings-acl.248, https://aclanthology.org/2021.findings-acl.248
16. Mei, Q., Shen, X., Zhai, C.: Automatic labeling of multinomial topic models. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 490–499 (2007)
17. Mueller, H., Rauh, C.: Reading between the lines: Prediction of political violence using newspaper text. American Political Science Review **112**(2), 358–375 (2018)
18. Popa, C., Rebedea, T.: BART-TL: Weakly-supervised topic label generation. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 1418–1425 (2021)
19. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3982–3992 (2019)
20. STT: Finnish news agency archive 1992-2018, source (http://urn.fi/urn:nbn:fi:lb-2019041501) (2019)
21. STT, Helsingin yliopisto, Alnajjar, K.: Finnish News Agency Archive 1992-2018, CoNLL-U, source (http://urn.fi/urn:nbn:fi:lb-2020031201) (2020), http://urn.fi/urn:nbn:fi:lb-2020031201
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
23. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating Text Generation with BERT. In: International Conference on Learning Representations (2019)