BACHELOR

Multilingual topic modelling on news data

Maksimovic, Nastasija

Award date:
2022

Technische Universiteit Eindhoven
University of Technology

Department of Mathematics and Computer Science

Multilingual topic modelling on news data

Nastasija Maksimovic (1524496)

Supervisors:
Uzay Kaymak
Emil Rijcken

Eindhoven, March 2022

Contents

Chapter 1

Introduction

Nowadays, news quality is becoming overwhelmingly important when informing the public about events and crucial topics in the everyday world. As seen with the notion of COVID19, some issues and reported events connected to them often gain more significance with the reading population, depending on the subjective perception of their relevance. Moreover, there is a demand for articles on specific topics, as readers become more anxious to get the latest and most important information. News agencies are becoming more focused on employing data analysis and machine learning to analyze customer engagement and news topic modelling to advance their businesses further and provide the highest quality content. One of the companies on a mission to make this possible and practical for newsrooms is the Dutch-Serbian merger company Smartocto. They deliver an intelligent editorial real-time analytics system to their clients to optimize storytellers' output and make it more relevant, impactful, and profitable. With the developments in Natural Language Processing, hereinafter referred to as NLP, Smartocto and their clients are advancing the idea of intelligent and relevant topic modelling in news articles to organize knowledge with different content better. As the company has clients all over the globe, the need to adapt methods to multilingual settings is of high importance to them. This thesis focuses on finding, understanding and implementing methods beneficial to the company, given the physical, social, and experimental context limitations discussed in section 1.3.

1.1 Problem definition

Various statistical methods engage in solving the objective of topic extraction. However, extracting topics across multilingual data is a relatively new subject in NLP. To build a topic model in a specific language, there is a need for a significant amount of training documents in that language, which is not always easy to find, especially for low-resource languages. Additionally, multilingual topic extraction tasks are usually researched on a single source, data, with parallel documents (documents with the same content in different languages), such as Wikipedia documents. Not much research has been done on the so-called real-life data, such as news data from various newsrooms, authors, countries of origin and hence different languages. Finally, most topic extraction methods generally require re-deriving inference networks due to any model changes, including adding new data, which is very time costly and inefficient for real-time analytics. This research aims to evaluate different ways to model real multilingual textual data from various news agencies worldwide without creating a separate model for each language and therefore supporting topic extraction on low resource languages. Additionally, this research places emphasis on models that can provide predictions fast, without re-deriving inference for each prediction. This will help Smartocto deliver better analysis and recommendations on customer engagement for specific news topics to their international clients seeking data analysis in different languages. As the company develops its editorial real-time analytics system, a multilingual topic extraction model would allow them to have better solutions for clients who deliver news in languages that are not widely used, as well as save them time on training new models for documents in languages they have not used before . The research is done in collaboration with Paula Dodig, who conducted the sister study on ”Evaluation and comparison of diverse topic models for English news data” (Paula Dodig, 2022).

1.2 Research question

Therefore, the following is considered as the main research question:

Which method is the most suitable for multilingual topic modelling for newspaper analytics?

1.2.1 Sub-questions:

- Which methods of multilingual topic modelling are available, and how to select the one(s) to implement given the solution criteria?
- How do the chosen models work and how to implement them?
- What is a good approach to evaluate the chosen methods?
- How well can the model predict topics for different languages according to the selected evaluation measures?

Answering these sub-questions will provide us with the required information to answer the main research question and provide the company Smartocoto with the most appropriate method needed for their purposes.

1.3 General settings and Scope

The scope of this research is limited by the restrictions and responsibilities defined in this section of the thesis.

- The data used to train this model is limited to news articles provided by the Smartocto’s clients, having that the research is done in collaboration with the company.
- The research is limited to the methods selected during the literature overview.
- Only methods implementable within Python libraries were considered for this research, constrained by the thesis author’s knowledge and capabilities.
- As a practical limitation, all computations needed for this research must be executable on a personal computer within a twenty-four-hour period. However, since the research is dedicated to the real-time analytics company Smartocto, the efficiency of the methods is highly valued.
- The research is limited to three languages (English, German and Dutch), on account of the large amounts of data provided by Smartocto in these languages.
- The topic distributions are to be modelled on medium-sized documents (medium-sized documents are usually around 1000 words with no strict constraints), having that we are dealing with news data.

1.4 Thesis outline

In the following chapters, exact detail and results of this research are provided. The chapter 2 on literature overview provides the necessary background information on multilingual topic modelling with a focus on efficient models useful for real-time analytics, existing methods and approaches as well as some other information needed to steer all the steps of this project. The reader gets introduced to the problem in more detail and gains knowledge of the current state of the academic community. In the methods chapter 3, specific methods and approaches leading to the answers to research questions are outlined and explained. There is more detail on the model selection process, data handling, exact formulas and algorithms. In data and experimental design, chapter 4, we give specifics of the experimental process. Following the methodology, this experimental setup explains which specific parameters are chosen and how the results are obtained to answer the questions. Following is the chapter 5 on the results and their interpretation - which methods perform well given our methodology and evaluation scheme and how. Lastly, there is the conclusion chapter 6 opens more questions that could follow from this research, mentioning its limitations and possible changes. The thesis is concluded with an overall deduction of insights and closure on the process.

Chapter 2

Literature Overview

2.1 Literature overview

An extensive literature search has been done to gather knowledge on the approaches already taken for similar problems and to understand multilingual topic modelling as a subject in general. The research has explicitly focused on methods available within the general settings and scope described in the section 1.3 of this article. Provided that Smartocto is a real-time analytics company, in the literature search, special attention was paid to methods that can efficiently approach the objective of multilingual topic extraction.[1]

The methods to model topics across multilingual corpora could be divided into three main subgroups (Lind et al., 2019):

- dividing the multilingual corpora into multiple monolingual corpora and creating a model for each language separately;
- creating one monolingual model and using it to extract topics in different languages by first translating them to the monolingual model training language using machine translation;
- using multilingual embeddings to create multilingual topic models, hereinafter referred to as MLTM, that can predict topics across different languages.

The first subgroup mentioned above has a lot of limitations. Traditional approaches to topic modelling are limited to one language, and sensitive to changes. In other words, they cannot use the knowledge gained on newly introduced data (transfer learning). Therefore, the main limitations of these models are that they cannot handle unknown words and hence cannot be applied to other languages because the vocabulary would not match. On the other hand, training a model on multiple languages introduces complexity issues such as the need for more data, more parameters and therefore slower training and possible overfitting (Boyd-Graber and Blei, 2009). According to the research on Autoencoding Variational Inference for topic models (Srivastava and Sutton, 2017), the above-mentioned limitations can be overcome by using neural variational autoencoders to approximate Latent Dirichlet Allocation. A method implementing variational autoencoders on topic modelling is called Neural PROD-LDA. To better understand the functioning of this model the underlying part of the model will be briefly explained in the subsection 3.1.1 of this research.

To get an overview of possible approaches for the second subgroup, research has been done to explore different machine translation methods. The two most widely known machine learning algorithms, Google Translate and DeepL Translator, were considered for this research. Both algorithms are based on machine-translation and use advanced artificial intelligence systems to translate natural languages. The biggest difference between the two competitor algorithms is the training datasets used to teach the algorithm how to solve the objective. While DeepL uses the

[1] Literature search conducted for this research did not focus on researching different topic extraction methods, given that this was covered by the sister study, conducted by (Paula Dodig, 2022).

online dictionary called Linguee, which contains large amounts of manual bilingual translations, Google Translate, uses multiple online sources of data, including the Europarl Corpus which is a set of multilingual content parallel documents manually translated to eleven languages containing the course of activities of the European Parliament in the time span of sixteen years. DeepL was proven to slightly outperform Google Translate at the task of machine translation according to research on evaluating machine translation methods, containing various quantitative and qualitative comparisons. More information on evaluation methodology and comparison of these two machine translation algorithms can be found in the following article (Isabelle et al., 2017).

Research has been done to explore different multilingual topic extraction models. Unlike monolingual models, multilingual topic models, do not have a separate representation of topics in different languages. Therefore, words with the same meaning in various languages have only one representation (Lind et al., 2019). This method is beneficial for modelling languages with little data in a specific domain. In order to have one topic representation for multiple languages, MLTMs need to solve the problem of aligning information across languages. There are two ways of modelling multilingual corpora or in other words, two types of information alignment (Lind et al., 2021):

- Lexical resources (The use of dictionaries to connect different languages).
- Multilingual embeddings (trained on topically comparable or parallel documents in multiple languages).

According to the research (Lind et al., 2021), the multilingual sentence embeddings are the most promising methods of solving the objective of multilingual topic extraction for the reasons that are further explained in the following section of this research. The two most widely known and used multilingual sentence embeddings are the LASER (language-agnostic sentence Representations) and the BERT's distiluse base multilingual sentence embedding (Yankovskaya et al., 2019). In addition to giving the topic extraction models multilingual properties, the above-mentioned models also account for contextual information, meaning the order of the words in the texts influences the model's decision, in addition to the text's content. LASER uses Bi-LSTM encoder-decoder architecture and machine translation to learn sentence representations (Artetxe and Schwenk, 2019). On the other side, BERT is based on transformer architecture. More specifically, BERT uses a bidirectional transformer encoder for learning word and sentence representations.

Finally, research has been done on the Zero-Shot topic model, which is an extension of the previously explained neural variational autoencoder model PROD-LDA. Instead of using a BOW representation of the input documents, Zero-Shot TM feeds the variational autoencoder with the SBERT multilingual embedding of the documents, providing the model with multilingual properties and allowing it to predict topics for languages the model is not trained on (Bianchi et al., 2021). More about this method will be discussed in the section 3.1.2.

Chapter 3

Methodology

3.1 Model selection

3.1.1 One Topic Model Per Language

This method approaches the problem of extracting topics on multilingual corpora by creating a separate topic model for each language. To implement this approach, the news article data was split into independent monolingual data sets (each data set contains only articles in one language). A monolingual model of choice is used as a baseline to compare the performance of the other two subgroups' models of choice. According to the results of the sister study (Paula Dodig, 2022), the model with the highest coherence score for topic modelling on news data is PROD-LDA. However, the research reported that the human-judgment evaluation of the topic coherence is higher for other methods that extract topics on monolingual news data evaluated in the same research, such as TopSBM. Nevertheless, the decision to use PROD-LDA was made for the following reasons. After further researching the models, promising insights into the potential of the PROD-LDA model were found. As it is based on variational autoencoders and therefore can handle missing words, avoids re-deriving inference, which allows for faster topic prediction and can be easily combined with multilingual embeddings, we chose to use it as our baseline model. Note that the model is only to be used for comparison purposes, given that for the monolingual corpora, other models have shown to outperform PROD-LDA.

To better understand the functioning of this model, the underlying part of the model will be briefly explained step by step.

LDA (Latent Dirichlet allocation)

LDA (Jelodar et al., 2019) is one of the most widely used topic models. It aims to discover the probability of each word in the document vocabulary being on a particular topic. It works on the principle of the following:

- Each document in our data (in each news article) is a mixture of topics.
- Each topic is a probability distribution over the vocabulary (the likelihood of a word belonging to a specific topic). The model learns how many documents are assigned to a particular topic because of a specific word. For more information (Jelodar et al., 2019).

The drawback of LDA is that it assumes that the documents can be represented with a Bag of Words representation, hereinafter referred to as BOW representation, which disregards the words' order and grammatical roles. This leads to some information loss, affecting the model's performance. Additionally, LDA needs to recalculate the inference network anytime changes are made to the topic model, making it highly inefficient.

Autoencoder

The encoder reduces the dimensionality of the data by producing a new compressed feature representation of the data and, in such a way, transforms it into a so-called latent space, or in other words, a hidden layer of a black box method. In contrast, the decoder aims to decompress the latent space. Autoencoders use neural networks to minimize the information loss in the encoder-decoder process of dimensionality reduction. This is done by training the bottleneck created by the encoder and the decoder using an iterative optimization process (gradient descent) to minimize the reconstruction error or lose as little information as possible during the encode-decode process. More information can be found here (Srivastava and Sutton, 2017). It is crucial to keep the depth of autoencoders controlled to avoid overfitting. Low reconstruction loss leads to a lack of regularity (interoperability and exploitability) in the data.

Variational Autoencoders (VAE)

Variational Autoencoders (VAE) (Srivastava and Sutton, 2017) are autoencoders whose encodings distribution is regularised during the training to ensure that its latent space has good properties allowing us to generate new data. Therefore, variational autoencoders regularise the training process to ensure that the latent space has good properties for the generative decoder process. This is done by encoding the input space as a Gaussian distribution over the latent space instead of as a vector, as it is done by Autoencoders. The encoder is trained to return the input distributions' mean and variance. By adding a variance component to the single point, we can regularise the latent space by enforcing it close to the standard normal distribution. The model still tries to minimize the information loss. However, we ensure that the latent space is well organized by adding the regularisation term. The regularisation term is the Kulback-Leibler divergence between the distributions returned by the encoder and the standard normal distribution. In the Figure 3.1, a visualization of a regularized (on the right) and non-regularized (on the left) latent space is shown. Different colours represent separate input distributions. When sampling from overlapping regions, the generated output is altered. It is essential to remember that by using variational autoencoders, we add an additional parameter to each data point, so the total number of model parameters doubles, influencing the model's execution time.

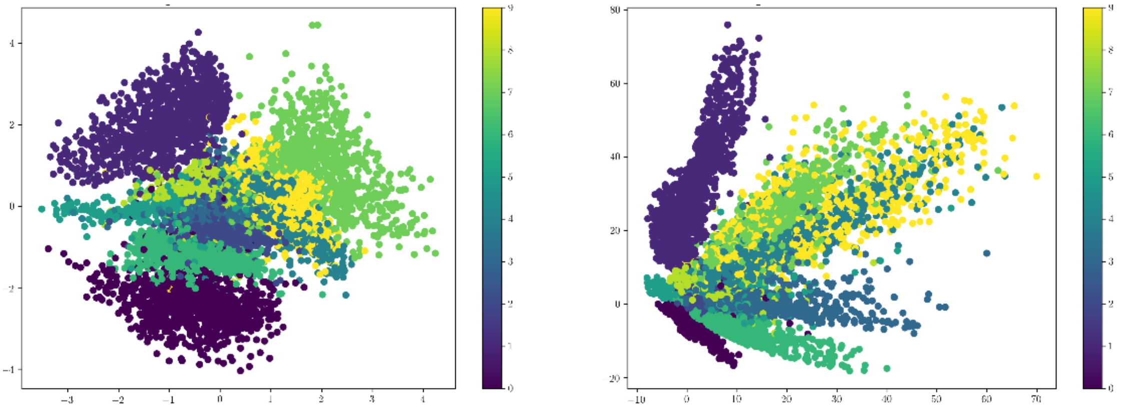


Figure 3.1: Latent space of Variational Autoencoder (Anwar, 2022)

PROD-LDA

VAE aim to teach a neural network how to mimic a posterior approximate inference network. Neural ProdLDA (Srivastava and Sutton, 2017)is a neural topic model based on variational autoencoders. To use the reparametrization trick(explained later in this section), PROD-LDA uses the Softmax-normal distribution, which is a location-scale family (family of distributions formed by translation and rescaling of a standard family member), to approximate the Dirichlet prior distribution. It is important for the prior distribution to be a location-scale family distribution in order for the back propagation to be applied to the distribution parameters. The neural variational framework takes as input a bag of words representation of the documents and trains the encoder network to approximate the mean and standard deviations parameters of the Gaussian distribution of document topics by compressing the input space. To estimate the expectation of the continuous latent variable (topic distribution), the model uses Monte Carlo sampling simulation and takes the weighted sum of the samples. Finally, the Monte Carlo estimate is used to optimize the variational parameters. The model uses the reparametrization trick to sample from the approximate posterior distribution, adding a fixed stochastic node randomly sampled from a standard normal distribution to add randomness to the sampling process. This allows the model parameters to be back-propagated by keeping them differentiable while still being able to sample from the approximate topic distribution. The optimization is done jointly on the generative model and inference network parameters with the Adam optimizer, which collectively, with added batch normalization and dropout units in the encoder network ensures that the encoder network does not get stuck in a bad local optimum and in such way avoids component collapsing. Subsequently, the Softplus function is used on the latent space obtained to get the topic-document distributions. Finally, the decoder reconstructs the original bag of words representation of the documents by generating words from the continuous latent representation(Srivastava and Sutton, 2017).

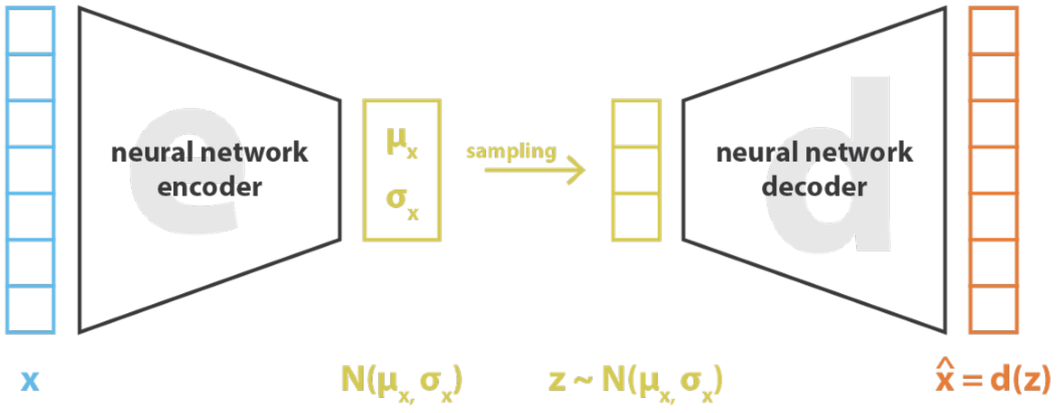


Figure 3.2: Variational Autoencoders Architecture (Rocca, 2021).

3.1.2 Document translation with Google Translate API

This method approaches the problem of extracting topics on multilingual corpora using machine translation (Lind et al., 2021). First, English was chosen as a baseline language (the language the model will be trained on). Using a machine translation algorithm, all non-English documents were translated to English. Both Google Translate API and DeepL Translator seem to make accurate translations most of the time (Yankovskaya et al., 2019). Therefore the selection was made based on the methodology scope of this research. Even though both algorithms were easily implementable in Python libraries, Google Translate offered an unlimited free API and a wider variability of languages translations can be done, which was the reason for choosing this algorithm over the DeepL Translator. In addition, Google Translate API offers the widest variety of languages, which benefits the company Smartocto, which provides clients all over the Globe with their services. The choice of the model implemented on the monolingual data set stays the same in the interest of comparability. Therefore, PROD-LDA was trained on the original English documents to extract topics, after which the predictions were made on the non-English documents translated to English. The evaluation metrics and the model's performance are discussed in the section 3.2 of this report.

Zero-Shot multilingual topic model

Smartoctos's clientele is very diverse; there is more data on some languages than others. Additionally, if the company were to get new clients with documents in new languages, it would be extremely beneficial if the model was able to extract topics on the new documents without taking the time and data resources to retrain the model. Hence, it is beneficial to the company if the topic extraction model can predict topics for unseen words and languages. Therefore the model chosen for this project is Zero-Shot TM. Zero-Shot TM is an extension of the previously explained baseline neural variational autoencoder model PROD-LDA. Instead of using a BOW representation of the input documents, Zero-Shot TM feeds the variational autoencoder with the SBERT multilingual embedding of the documents, providing the model with multilingual properties and allowing it to predict topics on languages the model is not trained on (Bianchi et al., 2021). Zero-Shot TM has three main advantages over the other considered models:

- Zero-Shot TM can handle missing words in the test set;
- Zero-Shot TM inherits the multilingual capabilities of recent pre-trained multilingual models;
- Zero-Shot TM does not require recalculating the inference network for each new prediction;

The Figure 3.3 shows the architecture of the previously mentioned Zero-Shot TM. The Contextualized embedding is the input data pre-trained on the multilingual BERT transformer model. The input is passed throw the hidden layer which compresses the documents into two parameter values of a Gaussian distribution. From the extracted distribution, the decoder re-samples a new reconstructed BOW representation of the documents, which is used to obtain the topics, or more specifically, most likely words that describe the topics.

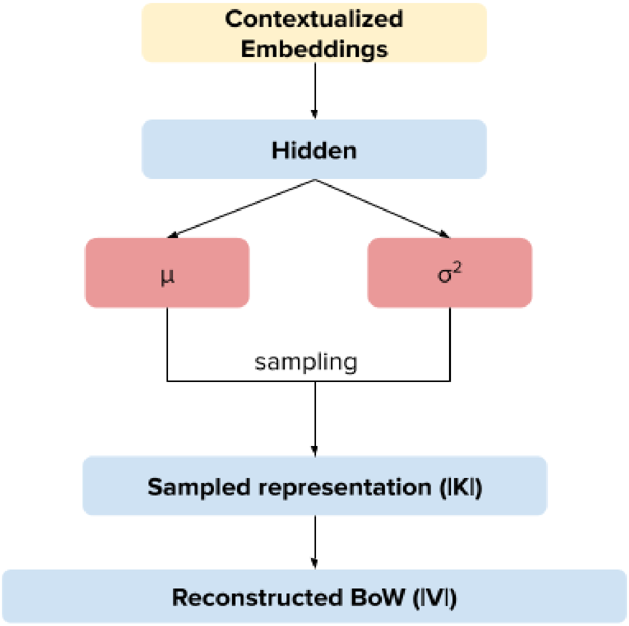


Figure 3.3: Zero-Shot Topic model architecture (Bianchi et al., 2021)

BERT embedding

The previously explained BERT multilingual sentence embedding (section 2.1), uses a masked language model to learn the bidirectional representation of the sentences and the next sentence prediction to learn which sentences are similar in meaning and go after each other. BERT's distiluse base multilingual sentence embedding is an improvement over the previous BERT models like mBERT (multilingual BERT) and sBERT (sentence BERT). Unlike BERT's distiluse base multilingual sentence embedding, mBERT does not align vectors of similar sentences in meaning across languages. Therefore this model is a multilingual extension of the sBERT approach that uses knowledge distillation to teach the model-aligned multilingual representations of words and sentences (Reimers and Gurevych, 2020). The way how the vector representations of similar sentences in meaning are aligned across languages is by using the teacher-student model methodology. The teacher model is trained on a source language. The student model is thought to mimic the teacher model's output for the same input documents as well as for the source language input documents translated to another language. In such a way the model gives the same output for parallel documents. The Figure 3.4 shows a visual representation of the teacher-student methodology previously explained, and how the approach pushes the student model's vector embeddings to be close the teacher model. The student model has the following properties:

- Vector spaces across languages are aligned (translated sentences are mapped to the exact location of the original sentence)
- Vector space properties are adopted from the teacher model to the new languages of the student model

This approach is called Multilingual knowledge distillation, as the student model distills the knowledge of the teacher (Reimers and Gurevych, 2020). By using this embedding, we are opening the door to multilingual topic extraction, having that the BERT embedding provides a multilingual representation of the documents. The benefit of using multilingual embeddings is predicting topics for low resource data. The student model is trained using the mean squared error loss function to minimize the difference between the teacher and student models' predictions. The specific embedding used is called "distiluse-base-multilingual-cased-v1" embedding (sBERT.net, 2020), allowing for topic extraction in over fifteen languages. For the scope of this research, the variety of languages is enough; however, to expand the model to a wider variety of languages, "distiluse-base-multilingual-cased-v2" embedding could be used for predictions in over fifty languages (Reimers and Gurevych, 2020).

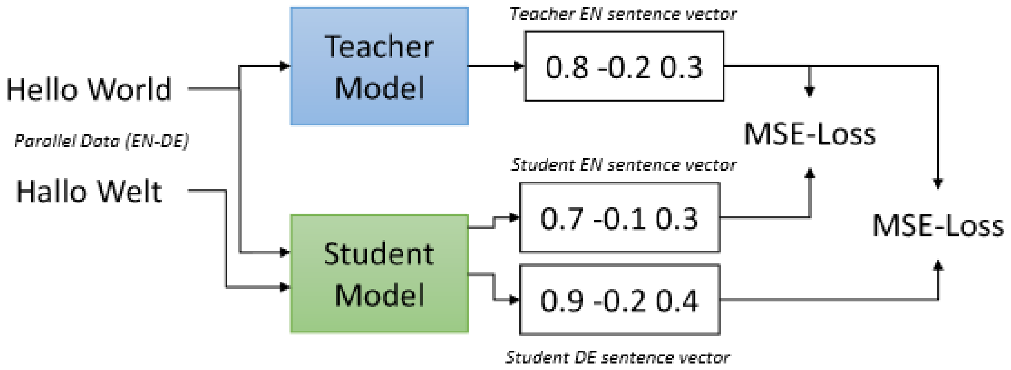


Figure 3.4: Teacher-Student architecture (Reimers and Gurevych, 2020)

3.2 Evaluation

As the comparison and evaluation stage will largely depend on the metrics and assessment criteria, we use monolingual PROD-LDA trained on English documents as the baseline model to which each multilingual model can be compared and evaluated against. In other words, PROD-LDA is to be used as a reference point for other models' evaluation. After running the metrics on each of the models, they will individually be compared to the PROD-LDA to determine the change in prediction quality when adding the multilingual properties to the baseline PROD-LDA model implemented. Further details of the experimental setup are discussed in the subsection 4.3.4 of this thesis.

To evaluate the models' predictions, we use automatic and manual assessment metrics:

1. Quantitative performance-based metrics.

2. Qualitative human-judgement-based metrics.

3.2.1 Quantitative performance-based metrics

Coherence

Topic Coherence score can be calculated to evaluate each topic by measuring the degree of semantic similarity between high-scoring words in the topic.

This measure, often used as a proxy for topic quality, is based on a hypothesis that words with similar meanings tend to co-occur in a similar (Syed and Spruit, 2017). There are multiple different coherence measures out of which we make use of the so-called C_V measure, found to have the highest correlation with human topic ranking data as found by Roder et al. (Röder et al., 2015). C_V is based on four parts (Syed and Spruit, 2017):

1. segmentation into parts where each of the topic's N words is paired with every other word;

2. calculation of word ($p(w_i)$) or word pair ($p(w_i, w_j)$) probabilities

3. calculation of a confirmation measure that quantifies how strongly a word set supports another word set using normalized pointwise mutual information (NMPI);

4. aggregation of individual confirmation measures;

In step three, words are represented as context vectors by:

$$\vec{v}(W') = [\sum_{w_i \in W'} NMPI(w_i, w_j)]_{j=1,\ldots|W|}$$

and the confirmation measure of a pair S_i as their cosine similarity:

$$\phi_{S_i}(\vec{u}, \vec{v}) = \frac{\sum_{i=1}^{|W|} u_i v_i}{||\vec{u}||_2 ||\vec{v}||_2}$$

Evaluation of the topic prediction on unseen languages

To evaluate how well the model predicts for unseen languages, it is necessary to have a parallel document (documents that have comparable content in different languages). Since this is not the case in our scenario, English documents were translated to German and Dutch and vice versa to compare the topic document distributions. This was done using the same methods as our translation model (Google translate API). By doing so we can calculate the following metrics for the model predicting in unseen languages:

- Kullback–Leibler Divergence measure (KL divergence)

  KL divergence quantifies how much one probability distribution differs from another by calculating the distributional similarity between the predicted topic distributions on the unseen language documents and the predicted topic distributions on the unseen language documents translated to the training language) (Bianchi et al., 2021).

$$KL(P||Q) = \sum_{x \in X} P(x) \ln \frac{P(x)}{Q(translated(x))}$$

  where $P(x)$ and $Q(translated(x))$ are the probability distributions compared.

- Centroid Distance measure

  CD computes the centroid embeddings of the top 10 words describing topics for the prediction on unseen language documents and for the prediction on unseen language documents translated to the training language, and then computes the cosine similarity for the corresponding two centroids. This metric is useful because it accounts for similar but not exactly the same topic predictions (Bianchi et al., 2021). The following is the formula for calculating the centroid embedding of the topics words that are the most likely to describe it:

$$C = \frac{sum(vector\_list)}{||(vector\_list||)}$$

  , where the vector list is the list of vector representations of the words that are most likely to describe a topic.

  The following formula is the cosine similarity between two centroid embeddings:

$$CD = 1 - \frac{C1 \cdot C2}{||C1||||C2||}$$

  , where C1 and C2 represent two centroid embeddings of topics.

3.2.2 Qualitative human judgment-based metrics

Topic prediction evaluation

Manual evaluation is chosen to rate how well the model predicts topics. Fifty random topic predictions will be rated manually on a scale of one to three, with three being the topic predicted is correct and one being the topic is entirely wrong. Then the average grade of each multilingual model will be compared to the baseline PROD-LDA model to see if adding multilingual properties to the model has changed its prediction accuracy and if the model can predict equally well on unseen languages as it can on the training language.

Topic interpretability evaluation

Manual evaluation will be used to assess the interpretability of the topics. Each model's topics will be assessed on the same scale (scale of one to three, with three being the topic predicted is correct and one being the topic is entirely wrong) to evaluate how interpretable the topics model has extracted are.

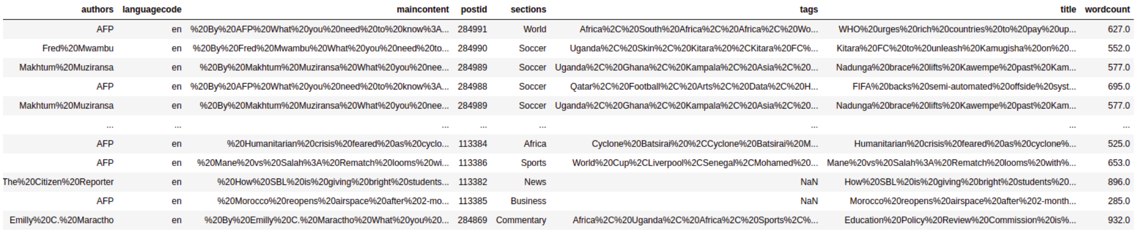
Chapter 4

Experimental Design

4.1 Data Description

The data in its raw form consisted of 738 csv files of variable size, altogether containing a one month's worth amount of articles from various clients of Smartocto. Articles published between the 7th of February and 9th of March 2022 had been crawled from the various websites and contain news in various languages about a wide variety of topics in countries all over the world, including sports, politics, culture and many others. Each csv file contains the following list of columns:

- atee_enabled (variable determining if for topic extraction will be provided to the client of that domain)
- authors (client name)
- create_date (date when the article was created)
- create_time (time when the article was created)
- domainid (id of the domain where it was published)
- error (error message if it arrises)
- imageurl (url of the images used in the articled)
- languagecode (two-letter code of the language in which the text was written)
- maincontent (main text with encoded special characters)
- maincontent_html (main text encoded in html)
- pid (text identificator from the client)
- postid (text identificator from the company)
- pubdate (date and time of publishing provided by the client)
- sections (encoded list of sections separated by a comma)
- tags (encoded list of tags separated by a comma)
- title (title of the text)
- url (the url address of the article)
- wordcount (number of words in the text)



authors	languagecode	maincontent	postid	sections	tags	title	wordcount
AFP	en	%20By%20AFP%20What%20you%20need%20to%20know%3A...	284991	World	Africa%2C%20South%20Africa%2C%20Africa%2C%20Wo...	WHO%20urges%20rich%20countries%20to%20pay%20up...	627.0
Fred%20Mwambu	en	%20By%20Fred%20Mwambu%20What%20you%20need%20to...	284990	Soccer	Uganda%2C%20Skin%2C%20Kitara%20%2CKitara%20FC%...	Kitara%20FC%20to%20unleash%20Kamugisha%20on%20...	552.0
Makhtum%20Muziransa	en	%20By%20Makhtum%20Muziransa%20What%20you%20nee...	284989	Soccer	Uganda%2C%20Ghana%2C%20Kampala%2C%20Asia%2C%20...	Nadunga%20brace%20lifts%20Kawempe%20past%20Kam...	577.0
AFP	en	%20By%20AFP%20What%20you%20need%20to%20know%3A...	284988	Soccer	Qatar%2C%20Football%2C%20Arts%2C%20Data%2C%20H...	FIFA%20backs%20semi-automated%20offside%20syst...	695.0
Makhtum%20Muziransa	en	%20By%20Makhtum%20Muziransa%20What%20you%20nee...	284989	Soccer	Uganda%2C%20Ghana%2C%20Kampala%2C%20Asia%2C%20...	Nadunga%20brace%20lifts%20Kawempe%20past%20Kam...	577.0
...	...	...	...	...	...	...	...
AFP	en	%20Humanitarian%20crisis%20feared%20as%20cyclo...	113384	Africa	Cyclone%20Batsirai%20%2CCyclone%20Batsirai%20M...	Humanitarian%20crisis%20feared%20as%20cyclone%...	525.0
AFP	en	%20Mane%20vs%20Salah%3A%20Rematch%20looms%20wi...	113386	Sports	World%20Cup%2CLiverpool%2CSenegal%2CMohamed%20...	Mane%20vs%20Salah%3A%20Rematch%20looms%20with%...	653.0
The%20Citizen%20Reporter	en	%20How%20SBL%20is%20giving%20bright%20students...	113382	News	NaN	How%20SBL%20is%20giving%20bright%20students%20...	896.0
AFP	en	%20Morocco%20reopens%20airspace%20after%202-mo...	113385	Business	NaN	Morocco%20reopens%20airspace%20after%202-month...	285.0
Emilly%20C.%20Maractho	en	%20By%20Emilly%20C.%20Maractho%20What%20you%20...	284869	Commentary	Africa%2C%20Uganda%2C%20Africa%2C%20Sports%2C%...	Education%20Policy%20Review%20Commission%20is%...	932.0

Figure 4.1: A snippet of the data table

The overall example table is available in figure 4.1.
To be able to utilize this amount of information in the best way possible, some filtering was needed. After all the files have been loaded into one unified DataFrame in Python, we were left with 57917 article entries. Many of the available columns were discarded since most topic modelling algorithms make little to no use of them in trying to extract and distribute the topics. Therefore, the text itself is the main and most important column that will be used in the whole project. Moreover, since we are focusing on the evaluation on only three out of seven provided languages, the data was filtered on the language code column. Considering that any missing values in the column with the main text would make the whole entry futile, all such entries were scrapped and deleted from the dataset. The columns that were kept were the following:

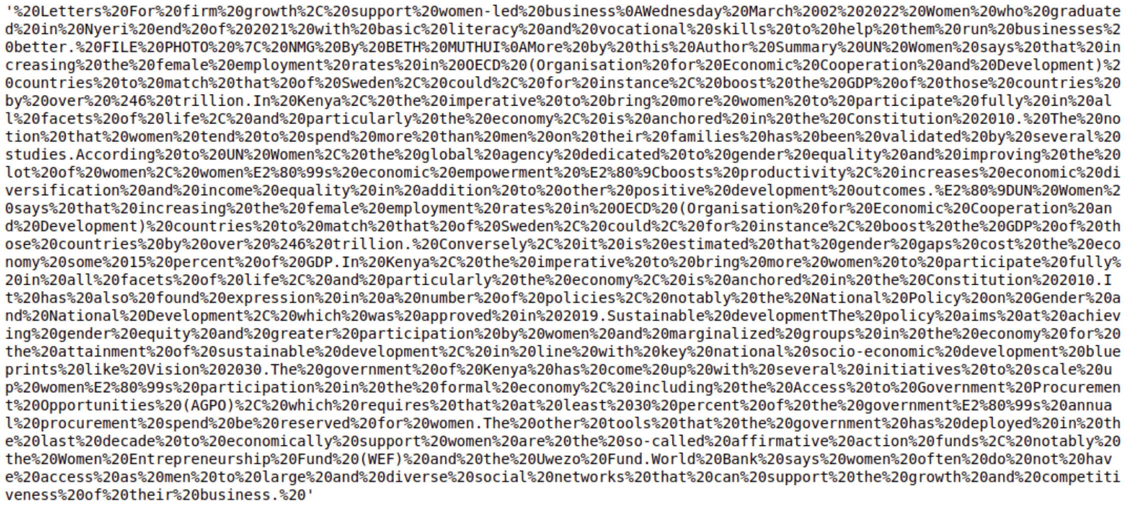
- authors
- languagecode
- maincontent
- postid
- sections
- tags
- title
- wordcount

The final result of the feasible data was 3378 articles in English, 2632 articles in German and 1860 articles in Dutch without missing values. Even though the most used column is actually just the maincontent containing the texts, the other columns serve as a comparison and easier indexing when it comes to extracting examples from the dataset.

4.2 Data Pre-processing

From here on, the pre-processing data stage takes place. The main text in its raw form can be seen in figure 5.4. A special character including the space character is encoded in the hexadecimal translation of ASCII code for punctuation. Some of the most prominent examples include $\%20, \%2C, \%22$ being a space, comma and colon. All the codes are to be translated into single character representation to determine if their use would contribute to the overall quality of the corpora. As this is mostly not the case with such characters, they were discarded and deleted from the text so that only the main words would be fed to the algorithms.

After a more detailed inspection of the articles, it was noticed that plenty of the data contains HTML buttons for social media and subscription messages at the end of the article, along with others. With the assumption that these primarily appear at the end of the page, there was no

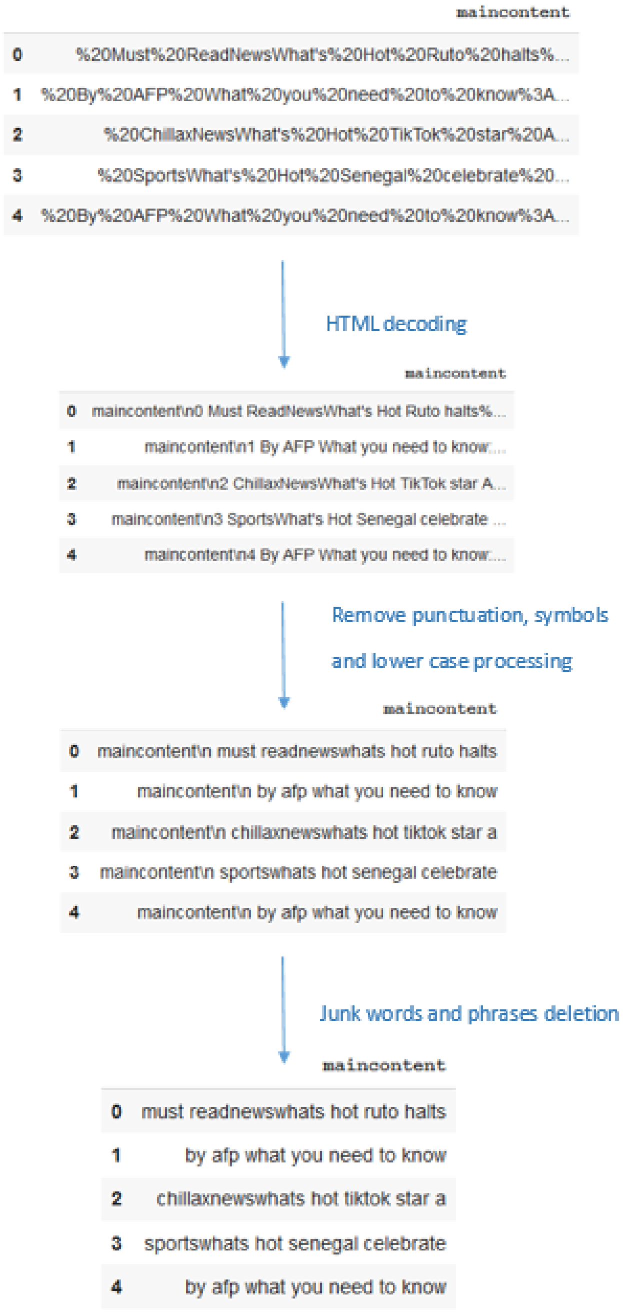


```
'%20Letters%20For%20firm%20growth%2C%20support%20women-led%20business%0AWednesday%20March%2002%202022%20Women%20who%20graduate
d%20in%20Nyeri%20end%20of%202021%20with%20basic%20literacy%20and%20vocational%20skills%20to%20help%20them%20run%20businesses%2
0better.%20FILE%20PHOTO%20%7C%20NMG%20By%20BETH%20MUTHUI%0AMore%20by%20this%20Author%20Summary%20UN%20Women%20says%20that%20in
creasing%20the%20female%20employment%20rates%20in%20OECD%20(Organisation%20for%20Economic%20Cooperation%20and%20Development)%2
0countries%20to%20match%20that%20of%20Sweden%2C%20could%2C%20for%20instance%2C%20boost%20the%20GDP%20of%20those%20countries%20
by%20over%20%246%20trillion.In%20Kenya%2C%20the%20imperative%20to%20bring%20more%20women%20to%20participate%20fully%20in%20al
l%20facets%20of%20life%2C%20and%20particularly%20the%20economy%2C%20is%20anchored%20in%20the%20Constitution%202010.%20The%20no
tion%20that%20women%20tend%20to%20spend%20more%20than%20men%20on%20their%20families%20has%20been%20validated%20by%20several%20
studies.According%20to%20UN%20Women%2C%20the%20global%20agency%20dedicated%20to%20gender%20equality%20and%20improving%20the%20
lot%20of%20women%2C%20women%E2%80%99s%20economic%20empowerment%20%E2%80%9Cboosts%20productivity%2C%20increases%20economic%20di
versification%20and%20income%20equality%20in%20addition%20to%20other%20positive%20development%20outcomes.%E2%80%9DUN%20Women%2
0says%20that%20increasing%20the%20female%20employment%20rates%20in%20OECD%20(Organisation%20for%20Economic%20Cooperation%20an
d%20Development)%20countries%20to%20match%20that%20of%20Sweden%2C%20could%2C%20for%20instance%2C%20boost%20the%20GDP%20of%20th
ose%20countries%20by%20over%20%246%20trillion.%20Conversely%2C%20it%20is%20estimated%20that%20gender%20gaps%20cost%20the%20eco
nomy%20some%2015%20percent%20of%20GDP.In%20Kenya%2C%20the%20imperative%20to%20bring%20more%20women%20to%20participate%20fully%
20in%20all%20facets%20of%20life%2C%20and%20particularly%20the%20economy%2C%20is%20anchored%20in%20the%20Constitution%202010.I
t%20has%20also%20found%20expression%20in%20a%20number%20of%20policies%2C%20notably%20the%20National%20Policy%20on%20Gender%20a
nd%20National%20Development%2C%20which%20was%20approved%20in%202019.Sustainable%20developmentThe%20policy%20aims%20at%20achiev
ing%20gender%20equity%20and%20greater%20participation%20by%20women%20and%20marginalized%20groups%20in%20the%20economy%20for%20
the%20attainment%20of%20sustainable%20development%2C%20in%20line%20with%20key%20national%20socio-economic%20development%20blue
prints%20like%20Vision%202030.The%20government%20of%20Kenya%20has%20come%20up%20with%20several%20initiatives%20to%20scale%20u
p%20women%E2%80%99s%20participation%20in%20the%20formal%20economy%2C%20including%20the%20Access%20to%20Government%20Procuremen
t%20Opportunities%20(AGPO)%2C%20which%20requires%20that%20at%20least%2030%20percent%20of%20the%20government%E2%80%99s%20annua
l%20procurement%20spend%20be%20reserved%20for%20women.The%20other%20tools%20that%20the%20government%20has%20deployed%20in%20th
e%20last%20decade%20to%20economically%20support%20women%20are%20the%20so-called%20affirmative%20action%20funds%2C%20notably%20
the%20Women%20Entrepreneurship%20Fund%20(WEF)%20and%20the%20Uwezo%20Fund.World%20Bank%20says%20women%20often%20do%20not%20hav
e%20access%20as%20men%20to%20large%20and%20diverse%20social%20networks%20that%20can%20support%20the%20growth%20and%20competiti
veness%20of%20their%20business.%20'
```

Figure 4.2: Raw example

other way but to assess article endings to notice some reoccurring overhead text manually. As the articles crawled from the same agency's website often have the same messages, we copy the respective text that should be deleted and iteratively remove it from the dataset. The most recurrent example of this is the *"monitor empower uganda we come to you we are always looking for ways to improve our stories let us know what you liked and what we can improve on i've got feedback premium share"*. As these messages are sometimes intentionally varied expressly to make automatic deletion harder, there is little that can be done to identify all of such undesirable input. Therefore, it is still possible that there are articles with unwanted text compromising the quality of the results.

Subsequently, simple pre-processing was done on the data to filter out a few words, remove empty documents after training, and remove punctuation and stop-words. Given that the pre-processing was done to implement a sentence embedding topic model, which requires contextual information, we limited the amount of pre-processing done on the data. In addition to this pre-processing, for the baseline PROD-LDA model, the documents were transformed into lists of words. The same list documents were used to create the BERT embedding representation of the documents for the Zero-Shot topic model. In Figure 4.3 a snippet of data and how it has changed throw-out, the pre-processing steps are shown.



	maincontent
0	%20Must%20ReadNewsWhat's%20Hot%20Ruto%20halts%...
1	%20By%20AFP%20What%20you%20need%20to%20know%3A...
2	%20ChillaxNewsWhat's%20Hot%20TikTok%20star%20A...
3	%20SportsWhat's%20Hot%20Senegal%20celebrate%20...
4	%20By%20AFP%20What%20you%20need%20to%20know%3A...

	maincontent
0	maincontent\n0 Must ReadNewsWhat's Hot Ruto halts%...
1	maincontent\n1 By AFP What you need to know:...
2	maincontent\n2 ChillaxNewsWhat's Hot TikTok star A...
3	maincontent\n3 SportsWhat's Hot Senegal celebrate ...
4	maincontent\n4 By AFP What you need to know:...

	maincontent
0	maincontent\n must readnewswhats hot ruto halts
1	maincontent\n by afp what you need to know
2	maincontent\n chillaxnewswhats hot tiktok star a
3	maincontent\n sportswhats hot senegal celebrate
4	maincontent\n by afp what you need to know

	maincontent
0	must readnewswhats hot ruto halts
1	by afp what you need to know
2	chillaxnewswhats hot tiktok star a
3	sportswhats hot senegal celebrate
4	by afp what you need to know

Figure 4.3: Pre-processing steps

4.3 Experimental Setup

The experiment is designed to compare the quality of predictions of methods predicting on multilingual corpora compared to the baseline PROD-LDA model trained on English documents. This is done by first implementing the baseline PROD-LDA model, with which the other two models are to be compared with. Secondly, non-English documents (documents in German and Dutch) are translated to English and fed into the baseline PROD-LDA model to extract topics from them. Finally, a Zero-Shot multilingual topic model (an extension of the baseline POD-LDA using multilingual sentence embeddings BERT) is trained on English documents and tested on German and Dutch documents to extract the predicted topics. Predictions of topic distributions for all three languages are evaluated by qualitative and quantitative measures to compare the different methods of multilingual topic extraction to each other, as well as to the monolingual baseline model, and in such way find the best solution to the task of multilingual topic extraction on news data.

4.3.1 PROD-LDA

To implement PROD-LDA, the packages Pyro and Torch were used. All the linear and softmax layers were added with Torch.nn module while we obtained the ELBO function from Pyro to build the encoder and decoder. The model parameter settings, including the number of topics extracted, followed the sister-study advice are the following. The model is trained for one-hundred epochs and optimized using the ADAM optimizer with a learning rate of 2e-3. A dropout of twenty per cent is applied to the document representations. The batch size used is 20, and the experiments are repeated twenty times (sample size is equal to twenty) and averaged out to provide a more stable prediction. According to the sister study results, the optimal number of topics to be extracted is twenty and sixty-five. This advice was followed for all models using PROD-LDA implemented in this research for comparison reasons.

4.3.2 Machine Translation

The chosen machine-translation method, Google Translate API, was used to translate all texts in German and Dutch to English to predict the topic of these documents using the English-trained baseline PROD-LDA model. Due to disfunction in the newer versions of the package, an older version (googletrans==3.1.0a0) was used to be able to use the algorithm iteratively.

4.3.3 Zero-Shot TM

The Zero-Shot topic model was implemented using the contextualized-topic-model python package provided by the research of Cross-lingual Contextualized Topic Models with Zero-shot Learning (Bianchi et al., 2021). The training set is created using both the un-pre-processed documents used for the contextual embedding and the pre-processed documents used to recreate the BOW representation of the most frequent two-thousand words needed to obtain the topics or, more specifically, most likely words representing a topic). The model is trained for hundred epochs and the same number of topics (20 and 65, respectively) as the baseline PROD-LDA model for comparison purposes. The Zero-Shot topic model is an implementation made available by (Bianchi et al., 2021). The parameter settings for the PROD-LDA model are the same as for the baseline PROD-LDA for comparison purposes.

Parameter settings for:	**PROD-LDA**	**Zero-Shot TM**
Library used for implementation	Pyro	contextualized-topic-models
Number of topics model is trained on	0.13	0.15
Number of epochs	100	100
Batch size	20	20
Optimization	Adam optimizer	Adam optimizer
Learning rate	2e-3	2e-3
Dropout rate	20%	20%

4.3.4 Evaluation

To calculate the KL divergence and centroid distance measures, predicted topic distributions for a set of two parallel documents in different languages. Having that, we are working with real-life data, and therefore no cross-language parallel documents are available to us; we created them by translating English documents to German and Dutch and predicting topics for the translations, to compare the difference in topic prediction and in such way test the multilingual property of the model. The translation was done using the same algorithm discussed in subsection 3.1.2.

Topic coherence is calculated for the baseline PROD-LDA model and for the Zero-Shot TM to investigate how sentence embedding influences the topic coherence.

Finally, manual evaluation is done on both the baseline PROD-LDA model and the Zero-Shot TM. For the Zero-Shot model, we evaluated both the prediction in the training language and the predictions in the unseen languages to compare it with the PROD-LDA. The comparison of the English prediction on the Zero-Shot TM with the baseline will allow us to see how does the sentence embedding and the contextual property of the BERT topic model influence the topic extraction, while the comparison of the prediction for the test documents in an unseen language will allow us to conclude if the model can predict as well on unseen languages as it can for the training language.

Figure 4.4 shows a visual representation of the experimental design of this thesis.

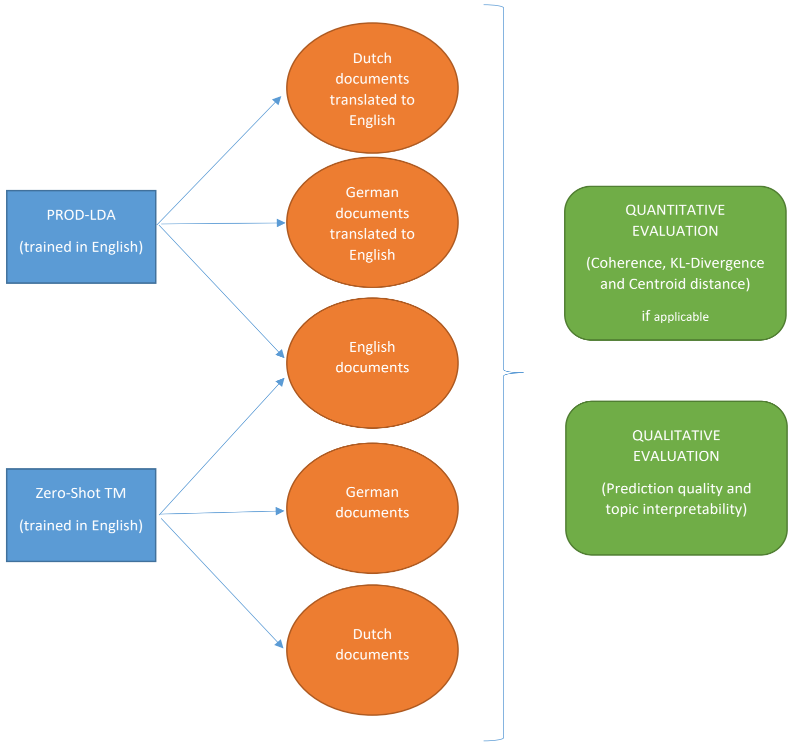

PROD-LDA
(trained in English)
Zero-Shot TM
(trained in English)
Dutch documents translated to English
German documents translated to English
English documents
German documents
Dutch documents
QUANTITATIVE EVALUATION
(Coherence, KL-Divergence and Centroid distance)
if applicable
QUALITATIVE EVALUATION
(Prediction quality and topic interpretability)

Figure 4.4: Visualization of the Experimental Design of this Thesis

Chapter 5

Results

5.1 PROD-LDA

- The coherence scores for the PROD-LDA model for 20 and 65 topics are 0.5 0.55 respectively Figure 5.3.
- The human evaluated topic interpretability average score for the 20 and 65 topics respectively are both equivalent to 0% intrepretability Figure 5.3.
- The human evaluation for the topic predictions accuracy was not done, having that the topic interpretability was so low that, it was hard to evaluate the prediction.

In the following figures 5.1 and 5.2, we show the first six topics the PROD-LDA model has extracted for 20 and for sixty-five topics respectively.

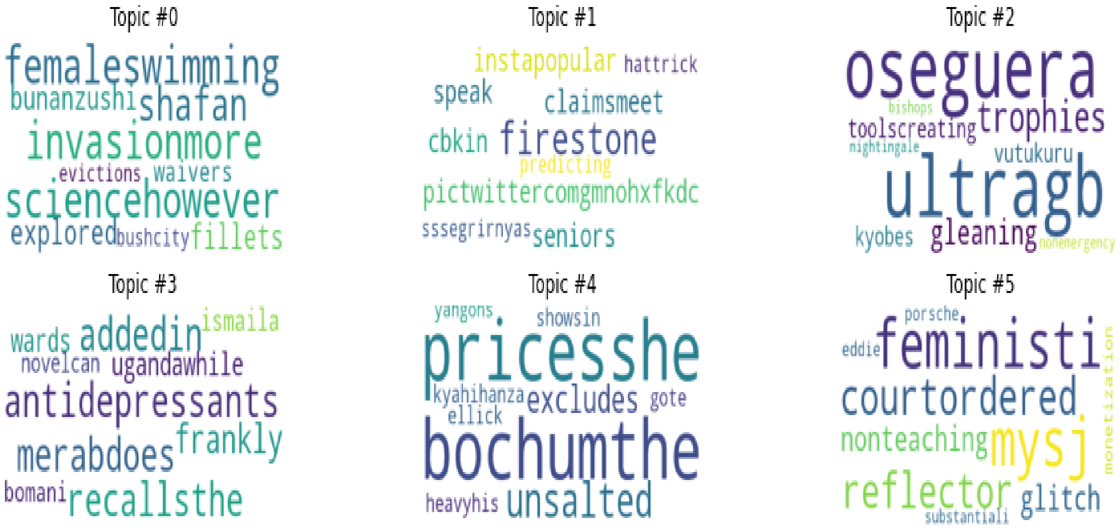


Figure 5.1: Topics extracted by PROD-LDA with 20 topics in English

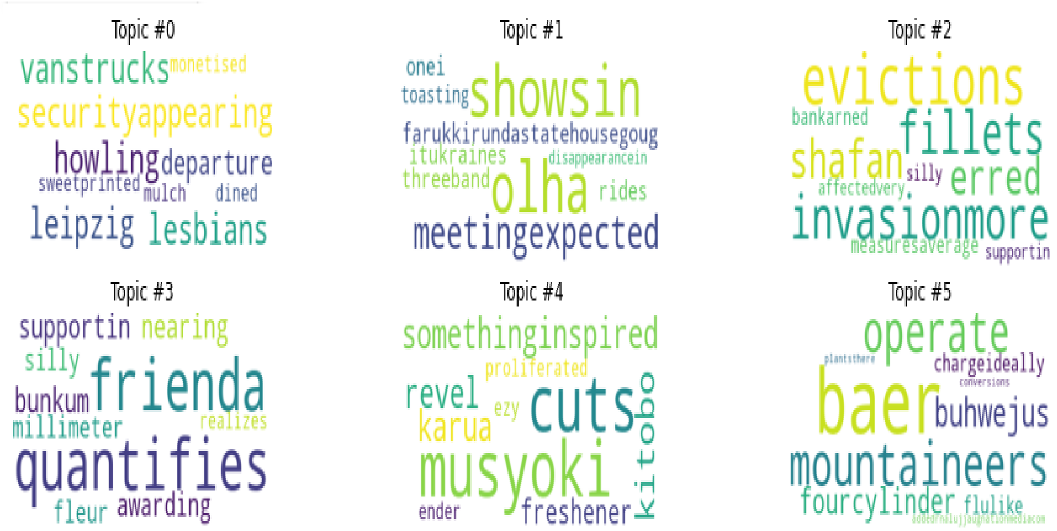


Figure 5.2: Topics extracted by PROD-LDA with sixty-five topics in English

5.2 Translation Model

- The human evaluated topic interpretability average score for the 20 and 65 topics respectively are 0.5 0.55 (same as for the baseline PROD-LDA model, having that the topics are not re-inferred).
- The human evaluation for the topic predictions accuracy was not done, having that the topic interpretability was so low that, it was hard to evaluate the prediction.

5.3 Zero-Shot Topic Model

- The coherence scores for the Zero-Shot model for 20 and 65 topics are 0.5 and 0.55 respectively Figure 5.3.
- The average KL-divergence score for the non-English topic predictions accuracy for 20 and 65 topics are 0.12 and 0.15 respectively Figure 5.3.
- The average Centroid similarity score for the non-English topic predictions accuracy for 20 and 65 topics are 0.68 and 0.58 respectively Figure 5.3.
- The human evaluated topic interpretability average scores for the 20 and 65 topics respectively are equivalent to 2.4 and 2.38. Figure 5.3
- The human evaluation scores for the English topic predictions accuracy for 20 and 65 topics are equivalent to 1.9 and 1.8 respectively Figure 5.3.
- The average human evaluation scores for the non-English topic predictions accuracy for 20 and 65 topics are equivalent to 1.45 and 1.35 respectively Figure 5.3.

The topics produced by Zero-Shot topic model are much more interpretable to a human. Although the topics are not perfect, most topics could be easily interpreted. For example:

- topic 0: Macroeconomics
- topic 1: Development of Uganda
- topic 3: Russia-Ukraine war

- topic 4: Kenya oil economy

In the following figures 5.3 and 5.4, we show the first six topics the Zero-Shot TM model has extracted for 20 and for sixty-five topics respectively:

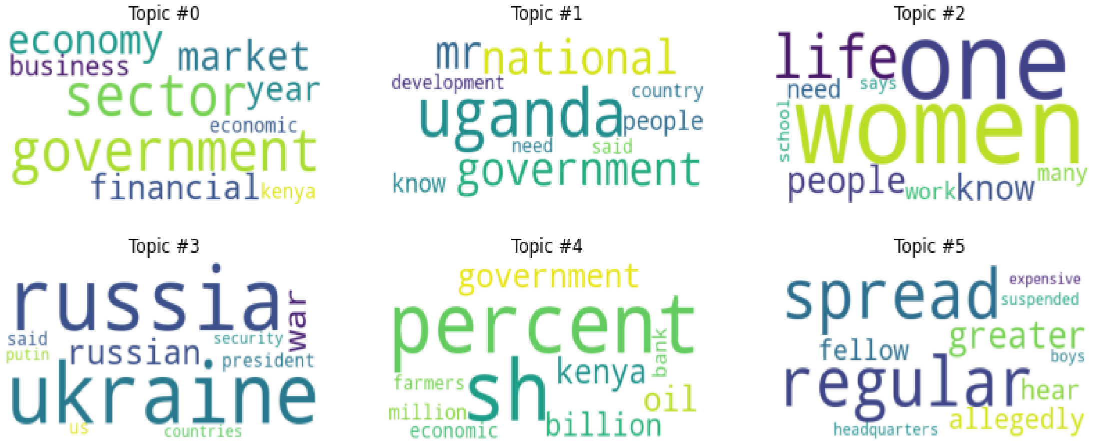


Figure 5.3: Topics extracted by Zero-Shot TM with 20 topics

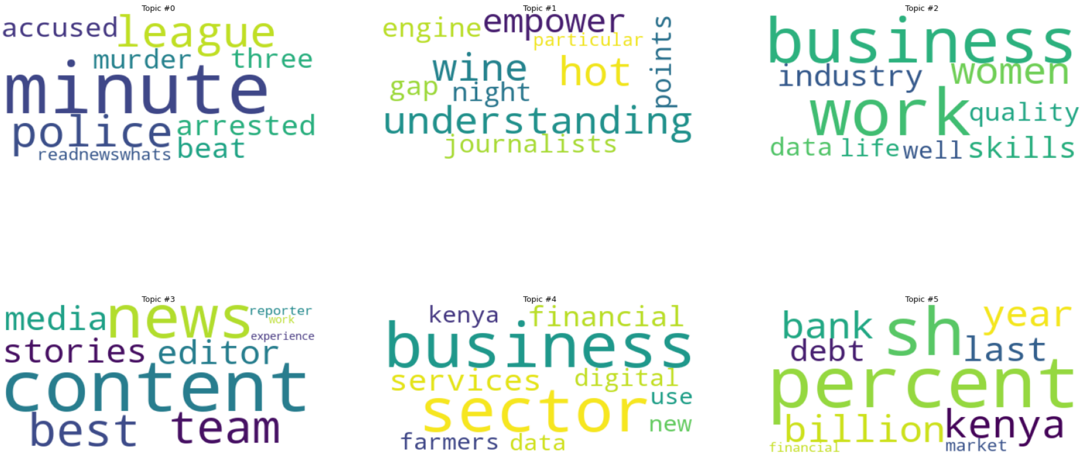
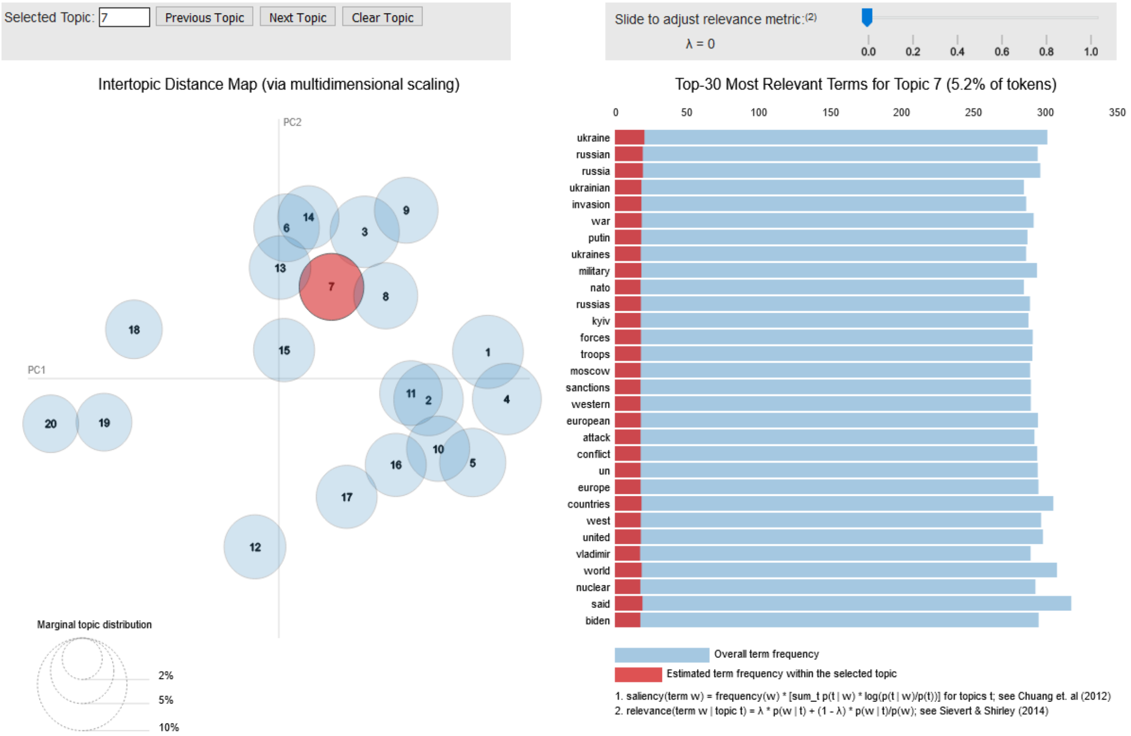


Figure 5.4: Topics extracted by Zero-Shot TM with sixty-five topics

Additional spatial representation of topics on a distance map was made and shown in Figure 5.5.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 5.5: Intertopic distance map representation of topics for 20 topics Zero-Shot model

CV Coherence score for:	**20 Topics**	**65 Topics**
PROD-LDA on English documents	0.5	0.55
Zero-Shot Topic Model on English documents	0.55	0.45

KL-divergence score for topic distributions for:	**20 Topics**	**65 Topics**
English documents and English documents translated to German English documents and English documents translated to Dutch	0.11 0.13	0.14 0.15
Dutch documents and Dutch documents translated to English	0.12	0.15
German documents and German documents translated to English	0.12	0.16

Centroid distance score for topic distributions for:	**20 Topics**	**65 Topics**
English documents and English documents translated to German	0.64	0.55
English documents and English documents translated to Dutch	0.77	0.69
German documents and German documents translated to English	0.62	0.52
Dutch documents and Dutch documents translated to English	0.68	0.58

Topic intrepretability average score (1-3) for:	**20 Topics**	**65 Topics**
PROD-LDA	1	1
Zero-Shot Topic Model	2.4	2.38

Human-judgment average score for:	20 Topics	65 Topics
PROD-LDA predictions for training documents	/	/
PROD-LDA predictions for test documents translated from German to English	/	/
PROD-LDA predictions for test documents translated from Dutch to English	/	/
Zero-Shot Topic Model predictions for training documents	1.9	1.8
Zero-Shot Topic Model predictions for test documents in German	1.4	1.4
Zero-Shot Topic Model predictions for test documents in Dutch	1.5	1.3

5.4 Runtime

The code was run on both the GPU and the CPU. section 5.4 shows the runtime for one epoch of both our Zero-Shot and Neural-ProdLDA for 25 and 50 topics on the HP Pavilion x360. Neural-ProdLDA is slightly faster than our ZeroShotTM, due to the additional time taken to train the BERT embedding. However, the run times are still comparable.

Run-time for one epoch:	One epoch for 20 Topics	One epoch for 65 Topics
PROD-LDA	1s	1s
Zero-Shot Topic Model	1.3s	1.3s

5.5 Discussion

The results shown in this chapter will be further summarized and interpreted in the following section. As can be seen from the section 5.1, even though the model was chosen for the reasons including having the highest coherence score out of the three other models it was compared to in the sister study (Paula Dodig, 2022), using only this measure is not a good indicator of the model's performance. After evaluating the interpretability of the topics extracted by the model, it was found that the keywords describing the topic were unrelated and, more importantly, morphed. As discussed in the chapter 3 and shown in the Figure 3.1, the importance of variational autoencoders is that by adding the KL-divergence term to the model, we ensure that the latent space is regularized and the amount of overlapping regions is minimized. The baseline PROD-LDA model seemingly did not do a good job at regularizing the latent space, which resolved in sampling from overlapping regions and therefore receiving a morphed reconstructed BOW representation of the documents from which the topics were constructed.

Since the translation approach of multilingual topic modelling was highly dependent on the baseline model's success, the approach could not be evaluated appropriately.

Finally, the Zero-Shot model coherence score was exactly the same for both 20 and 65 topic models. However, the human evaluation of the topics was significantly better, although still not perfect. Even though the topics' keywords are highly related to each other and could be interpreted into a broader concept, as shown in the example in section 5.3, during the topic interpretability evaluation, it was noticed that many topics had been repeated with some minor alterations to them. This was also visualized in Figure 5.5. The left side of the figure shows the visual representation of the topics in the latent space, whereas the right side shows the most relevant terms for the selected topic (in this case, topic 7). If the topics are similar, they are placed closer to each other, and topics distant in their meaning have a more significant distance between them in the graph. In the graph, we can see multiple clusters of topics. When the interactive graph was further explored, it was found that topics within the cluster were closely related. Increasing the number of topics did not resolve in receiving more diverse topics but rather more variations of the same topics. This may be because the number of articles on these reoccurring topics was so high that the model minimized the loss by always predicting these same topics. Therefore, it is assumed that the data the models are trained on should be carefully selected. The time frame in which the training articles are issued should be more than one month, which is not the case with our training data, to get a wider variety of topics. Additionally, to get a broader selection of topics, it would be beneficial to train the model on the data provided by a newsroom which are reporting on more general topics, rather than a newsroom with a particular audience, such as small country newsrooms reporting on local events or financial newsrooms which was the case with our training articles. This is especially important, having that the model has a predefined number of topics which are not re-inferred for later unseen documents' predictions. This explains the human evaluation scores for the topic predictions' accuracy. The large number of articles that were about those reoccurring topics have gotten an accurate topic prediction. However, there was still a great deal of articles about topics that were not even captured by the model and therefore have received a lousy topic prediction. As far as the predictions for documents in unseen languages go, the human evaluated accuracy of prediction decreased. During the human evaluation, it was noticed that, this was because the articles in unseen languages were about completely different subjects which could not be explained by the predefined topics. This is why, once again, it is emphasized that the training documents' selection is highly important. On the other hand, the KL-divergence scores for the translated English document predictions were fairly low (on average 0.13, whereas the value of 0 would indicate that the predicted topic distributions are exactly the same). Additionally, the centroid distance scores were fairly high (on average 0.63, whereas the value of 1 would mean that the centroid embedding of the predicted topics keywords for predictions on parallel documents in English and an unseen language are exactly the same). This tells us that, aside from the BERT embedding significantly increasing the topics' intrepretability, it is also making predictions on a similar level for the unseen documents as it is for the documents in the training language. Therefore, the biggest drawback of this model is that it cannot extract diverse predictions, but it captures the same topics multiple times, which might be fixed by training it on a better selection of training documents.

Chapter 6

Conclusion

6.1 Discussion and Further Research

Topic modelling, in general, is not an easy task. It works under the assumption that text data is generated from a specific distribution and tries to approximate it. The understanding of natural language knowledge is based on much pre-assumed knowledge of the reader, which is hard to translate to a model. In addition, meaning is highly dependent on the context of the text; therefore, disregarding the contextual information of text documents leads to a lot of information loss. As shown in this research, embedding data in a way that keeps the contextual information significantly improves the performance of the topic model.

Results found in this research give us an overview of how Variational autoencoders could be used to tackle the task of multilingual topic extraction on news data efficiently. The results suggest that, while using the neural variational autoencoder PROD-LDA solely on the data does not provide us with satisfactory results, by combining it with a multilingual sentence embedding BERT in a so called, Zero-Shot topic model offers a promising solution to the problem.

As discussed earlier in this article, the Zero-Shot TM extracts a specified number of predefined topics from the training data and assigns one of them to unseen articles, meaning it assumes that the gist of topics in the training documents is generalizable. Since we are dealing with news data, we expect the unseen articles to introduce new topics to the model, making it difficult to assign them to one of the pre-existing topics. Therefore, to keep the topics up to date, for further research it would be beneficial to re-execute the model regularly to keep the topics up to date. Considering that a considerable amount of news articles come out every day, it should not be difficult to gather enough news articles for the model to recognize them as a new topic. However, by doing so, we downsize the benefit of using the neural variational autoencoder since we still have to recalculate the inference quite frequently. Additionally, the variety of topics in the news is highly diverse; thus, establishing fixed topics and using them on unseen articles is not likely to produce the most accurate results and would require many topics. Subsequently, the more topics we want to define, the more data the model will need to learn them, which increases the execution time.

Alongside the challenge of aligning topics in different languages, the multilingual topic extraction task brings other difficulties. Namely, having articles in different languages usually means that they were written for different countries/regions and often report news concerning that specific region. If the news is not relevant globally, it is likely not reported in the other regions and languages. For instance, Dutch news may cover the Dutch National Cricket Championship, but German news will most likely not. So if we train our model in German, the topics we extract are fixed to the relevant and reported events in Germany. Hence, assigning one of the predefined topics to that document would be hard if we tried to predict topics for the Dutch Cricket article. Extracting the general article domain could help with this problem rather than topics that describe what a specific article is about. This could be done by reducing the specified number of topics to make the model learn more general long-term topics such as health, sports, investing etc. This approach was attempted in this research; however, when the model was trained to learn twenty topics or sixty-five topics, the amount of diversity within topics was still more or less the same. In other words, the model kept producing similar topics multiple times. This might be because the data we trained the model on was not very diverse. It would be beneficial to create a better training set which contains a wide diversity of non-country-specific topics (i.e. articles from an international newsroom).

Another reason may be that the data used to train the model were news articles published in the same month; hence, the model tried to capture trending topics rather than more general topics that reoccur in the news in the long term. This problem could be fixed by training the model on the news articles written over a more extended period of time. Having more data/articles is generally beneficial for the model's performance. In this case, the added benefit of training on larger corpora is that more news articles over a more extended period of time ensure that topics' diversity increases. Hence, with a low prespecified number of topics, the model would have to produce more general topics to fit each article to a topic. Additionally, by having fewer generalized topics, the problem of re-executing the model frequently to keep the topics up to date would be solved. However, this comes at the cost of having a more general insight into what the article is about rather than a specific topic. For further research it is proposed to make two models, one trained on carefully selected diverse documents, used to capture long term broad topics, while the other model is designed to capture the trending topics, and is trained on more specific data sets (i.e. data issued in a smaller time frame, or a particular region). Finally, the second model could also be a dynamic model to make sure it captures all trending topics (Blei and Lafferty, 2006).

I believe that the implemented Zero-Shot topic model has a lot of potential, and if further developed could be a very valuable tool for the Smartocto company. Further I have great hopes for the potential it holds for further multilingual topic modeling research.

Bibliography

Anwar, A. (2022). Difference between AutoEncoder (AE) and Variational AutoEncoder (VAE). 8

Artetxe, M. and Schwenk, H. (2019). Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7. 6

Bianchi, F., Terragni, S., Hovy, D., Nozza, D., and Fersini, E. (2021). Cross-lingual contextualized topic models with zero-shot learning. In *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference.* 6, 10, 13, 18

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *ACM International Conference Proceeding Series*, volume 148. 28

Boyd-Graber, J. and Blei, D. M. (2009). Multilingual topic models for unaligned text. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, UAI 2009.* 5

Hidalgo-Ternero, C. M. (2021). Google Translate vs. DeepL: analysing neural machine translation performance under the challenge of phraseological variation. *MonTI. Monographs in translation and interpreting*, pages 154–177.

Huber, J. and Spiliopoulou, M. (2019). Learning multilingual topics through aspect extraction from monolingual texts.

Isabelle, P., Cherry, C., and Foster, G. (2017). A challenge set approach to evaluating machine translation. In *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings.* 6

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11). 7

Lind, F., Eberl, J. M., Eisele, O., Heidenreich, T., Galyga, S., and Boomgaarden, H. G. (2021). Building the Bridge: Topic Modeling for Comparative Research. *Communication Methods and Measures.* 6, 9

Lind, F., Eberl, J.-M., Galyga, S., Heidenreich, T., Boomgaarden, H. G., Jiménez, B. H., and Berganza, R. (2019). A bridge over the language gap: Topic modelling for text analyses across languages for country comparative research. *University of Vienna: Working Paper of the REMINDER-Project.* 5, 6

Paula Dodig (2022). Evaluation and comparison of diverse topic models for English news data. 3, 5, 7, 25

Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference.* 11

Rocca, J. (2021). Understanding Variational Autoencoders (VAEs) - Towards Data Science. 9

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*. 12

sBERT.net (2020). SentenceTransformers Documentation — Sentence-Transformers documentation. 11

Srivastava, A. and Sutton, C. (2017). Autoencoding variational inference for topic models. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*. 5, 8, 9

Syed, S. and Spruit, M. (2017). Examining topic coherence scores using latent dirichlet allocation. In *Proceedings - 2017 International Conference on Data Science and Advanced Analytics, DSAA 2017*, volume 2018-January. 12

Tang, Y. K., Mao, X. L., Huang, H., Shi, X., and Wen, G. (2018). Conceptualization topic modeling. *Multimedia Tools and Applications*, 77(3).

Vulić, I., De Smet, W., Tang, J., and Moens, M. F. (2015). Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing and Management*, 51(1).

Xie, Q., Zhang, X., Ding, Y., and Song, M. (2020). Monolingual and multilingual topic analysis using LDA and BERT embeddings. *Journal of Informetrics*, 14(3).

Xue, J., Luo, J. J., Yuan, C., and Yamagata, T. (2020). Discovery of Chile Niño/Niña. *Geophysical Research Letters*, 47(5).

Yankovskaya, L., Tättar, A., and Fishel, M. (2019). Quality estimation and translation metrics via pre-trained word and sentence embeddings. In *WMT 2019 - 4th Conference on Machine Translation, Proceedings of the Conference*, volume 3. 6, 9

Yuan, M., Van Durme, B., and Boyd-Graber, J. (2018). Multilingual anchoring: Interactive topic modeling and alignment across languages. In *Advances in Neural Information Processing Systems*, volume 2018-December.

Zosa, E. and Granroth-Wilding, M. (2019). Multilingual dynamic topic model. In *International Conference Recent Advances in Natural Language Processing, RANLP*, volume 2019-September.