Kevin Bönisch
M.Sc. Computer Science
Specialized in Artificial Intelligence
k.boenisch@outlook.com

Beyond Topic Modelling: Introducing Topic Search with VecTop

Kevin Bönisch

11th March 2024

Text Technology Lab
Prof. Dr. Alexander Mehler

1 Introduction

The ability to automatically extract topics from texts, documents or collections is an ongoing task in the fields of artificial intelligence (AI) and natural language processing (NLP). Over the last few decades, a variety of modelling techniques have been created towards solving this endeavour, which can be broadly categorized into the disciplines of *Topic Modelling* and *Topic Classification* (which is also often referred to as *Text Classification*).
Topic modelling refers to the unsupervised task of extracting latent variables from large datasets (D. M. Blei 2012, p. 1), which is primarily suited for text data but also has its use cases in other disciplines such as bioinformatics (L. Liu et al. 2016) or environmental data exploration (Girdhar, Giguère, and Dudek 2013).
Topic classification alludes to the usage of supervised learning methodologies (Osnabrügge, Ash, and Morelli 2021, Q. Li et al. 2016), where presently, there exists a prevailing tendency towards employing pre-trained language models and transformers (Vaswani et al. 2023) for the purpose of fine-tuning in this domain (Peña et al. 2023, Sun et al. 2023, Z. Wang, Pang, and Lin 2023), since even without explicit domain-based fine-tuning, general instruction-following language models such as ChatGPT (OpenAI 2022), GPT-4 (OpenAI 2023), Vicuna (Chiang et al. 2023) or Stanford Alpaca (Taori et al. 2023) are capable of assigning texts with topics whilst offering a variety of different abilities as well.
However, since both strategies aim to automatically model and classify topics, criticism about their linguistic capabilities (Shadrova 2021, Schröter and Du 2022) and technical limitations (Laureate, Buntine, and Linger 2023, Barde and Bainwad 2017) are prominent.

Within this treatise, I will briefly layout the history of this discipline before illustrating frequent criticism and finally proposing a novel solution called VecTop, that aims to address some of the limitations of traditional and recent topic modelling and classification techniques. To do so, VecTop will utilize cross-bilingual word embeddings, k-nearest neighbors, extractive text summarization and a variety of corpora to classify unlabeled texts with a collection of topics and also subtopics. This process can neither be categorized as topic modelling nor topic classification, which is why I introduce the term *Topic Search* to describe VecTop's abilities. The source code for VecTop is open-source and available on GitHub[1].

[1] https://github.com/TheItCrOw/VecTop

2 Topic Modelling and Classification

In this section, I will showcase the current state of topic modelling and classification, beginning with the first efforts and progressing to the models that are employed today, as well as the issues they encounter. This section, besides the explicit citations in it, stems on the works of Churchill and Singh (2022), Kherwa and Bansal (2019) and Vayansky and Kumar (2020).

The earliest topic models date back to 1990, where Deerwester et al. (1990) described how latent semantic analysis can be used to automatically index and retrieve documents from large databases, which they called *Latent Semantic Indexing (LSI)*. LSI extracts topics through vectors of word frequencies that were derived using *singular value decomposition (SVD)*. Matching these vectors against the word frequency vectors of single documents, topics could be classified. This defines what would later become the widespread *bag-of-words* model. (Churchill and Singh 2022, p. 8).
Hofmann (1999) built on top of that by proposing the *Probabilistic Latent Semantic Indexing (pLSI)* which replaces the SVD with a generative *aspect model*. The aspect model represents documents and terms as mixtures of latent topics or *aspects*, where each document is assumed to be generated by a mixture of aspects and each aspect is on its own characterized through a probability distribution over terms. The key idea behind the aspect model is to model the generation of terms in a document through a generative process involving latent aspects. By estimating the parameters of this model, it is possible to perform tasks such as document retrieval, document classification and term clustering.
A year later, Nigam et al. (2000) explored the impact of incorporating unlableled data into text classification, again utilizing a generative model while also adding the *Dirichlet distribution* (see Section 2.1). This model would later be known as the *Dirichlet Multinomial Mixture (DMM)* model. By using unlabeled data with the addition of the Dirichlet distribution, the authors could improve results over labeled data by 30%, but the model falls short in other cases.

2.1 Latent Dirichlet Allocation

In 2001, D. Blei, Ng, and Jordan (2001) then proposed the widespread *Latent Dirichlet Allocation (LDA)* and therein coin the term *topic model*. LDA stems upon pLSI with the addition of the *Dirichlet distribution*. It still uses the equivalent bag-of-words model and document-term matrix from pLSI, but this time sampling a distribution of topics for a document instead of a single topic per document. The goal of LDA is to refine the distribution of topics across words in order to maximize the likelihood of documents within a dataset across a given number, k, of topics. Besides k, LDA utilizes the additional parameters α and β, where α is defined as the topics-per-document and β as the words-per-topic ratio. LDA's general process is outlined by algorithm 1.

LDA would spawn many variations over the years to come, such as the Hierarchical Dirichlet Processes (Teh et al. 2006), Correlated Topic Models (D. Blei and Lafferty 2006) or the *special words with background (SWB) model* (Chemudugunta, Smyth, and Steyvers 2006). The main problems identified by the authors at that time were the computational complexities of the proposed techniques, which were often NP-hard (Sontag and Roy 2009), and the lack of adapting to timely developments concerning topics as well as the documents themselves. Shadrova (2021) subsequently emphasized the deficiency in adapting to the evolving nature of language over time.t

```
Input:
  • Set of $M$ documents $D$

for $d \in D$ : do
    Randomly draw the number of words $N$ for $d$.
    Randomly draw the topic distribution $\theta$ from the Dirichlet distribution,
      conditioned on the parameter $\alpha$.
    for $w_i, 0 \leq i \leq N$ : do
        Draw a topic $z_i$ from $\theta$
        Draw a word $w_i$ based on the probability of $w_i$ given the topic $z_i$ and
          conditioned on the parameter $\beta$.
    end
end
```

Algorithm 1: LDA's process, modified from Churchill and Singh (2022, p. 10))

2.2 Towards Modern Topic Modelling and Classification

As was briefly alluded to, the change in vocabulary, documents and language has been far more drastic over the past years than the development of the models themselves. Slang, out-of-vocabulary words, and neologisms have grown in popularity; the format of documents has shifted from literary to shorter social media or blog writings, and the language itself has evolved more rapidly as a result. Despite the prompt change over the years, many of the old mathematical components from the decades prior are still used whilst also having spawned new approaches as I will showcase in the following.

2.2.1 Non-negative Matrix Factorization

Non-negative Matrix Factorization (NMF) is the mathematical equivalent to pLSI, given a certain error function (Ding, T. Li, and Peng 2008). Its premise stems from factorizing a non-negative matrix into two new matrices such that the product of those two is equal to the original, which is NP-hard (Vavasis 2010). Successfully applying NMF for topic modelling (Shahnaz et al. (2006), Yan et al. (2013)) is outlined by algorithm 2, where the matrices

W (document-topic matrix) and H (topic-word matrix) converge iteratively by minimizing the reconstruction error. After the convergence, the resulting matrices W and H represent the document-topic and topic-word distribution respectively. The intuition is to reconstruct through convergence which was factorized into W and H from the document-word matrix V, and hence represent each document as a mixture of topics.

```
Input  : Document-Word Matrix $V$ of shape $(m, n)$, where $m$ is the number of
         documents and $n$ is the number of words.
         Number of topics $k$
Output: Matrices $W$ (Topic-Word) and $H$ (Document-Topic)
Initialize $W$ with random non-negative values $(m, k)$;
Initialize $H$ with random non-negative values $(k, n)$;
while not converged do
    Update $W$:
    $W \leftarrow W \odot \left(\frac{V}{WH} H^\top\right)$ ;                // Element-wise multiplication
    Normalize columns of $W$ to sum to 1;
    Update $H$:
    $H \leftarrow H \odot \left(\frac{W^\top V}{W^\top WH}\right)$ ;          // Element-wise multiplication
    Normalize rows of $H$ to sum to 1;
end
return $W$, $H$
```

Algorithm 2: NMF (Lee and Seung 2000) adapted for Topic Modelling

2.2.2 Word Embedding Space

With the discovery that neural models could learn distributed representations for words (Bengio, Ducharme, and Vincent 2000), formerly called *word feature vectors*, Mikolov et al. (2013) were able to propose the *Word2Vec* model, that produces what we now know as *word embeddings*. These word vectors could be used to find semantically related words by projecting them into a vector space of fixed dimensionality, such that clusters of semantically similar words are created (see Figure 2.1). This endeavour in general is comparable to topic modelling and also the driving force behind VecTop as described in Section 3.

Consequently, Nguyen et al. (2015) used word embedding spaces in conjunction with the traditional topic models LDA and DMM by expanding the topic-word distribution with a latent feature component composed of word vectors. The results show improvement over traditional LDA, specifically on shorter texts.

Expanding the possibilities of word embedding spaces, Qiang et al. 2017 introduced an *embedding-based topic model (ETM)* using the Word2Vec framework, which holds similarities to VecTop. In it, they introduce the *Word Mover's Distance (WMD)*, which is a scale to measure the difference between documents given their Word2Vec vectors. Utilizing the WMD, they bundle

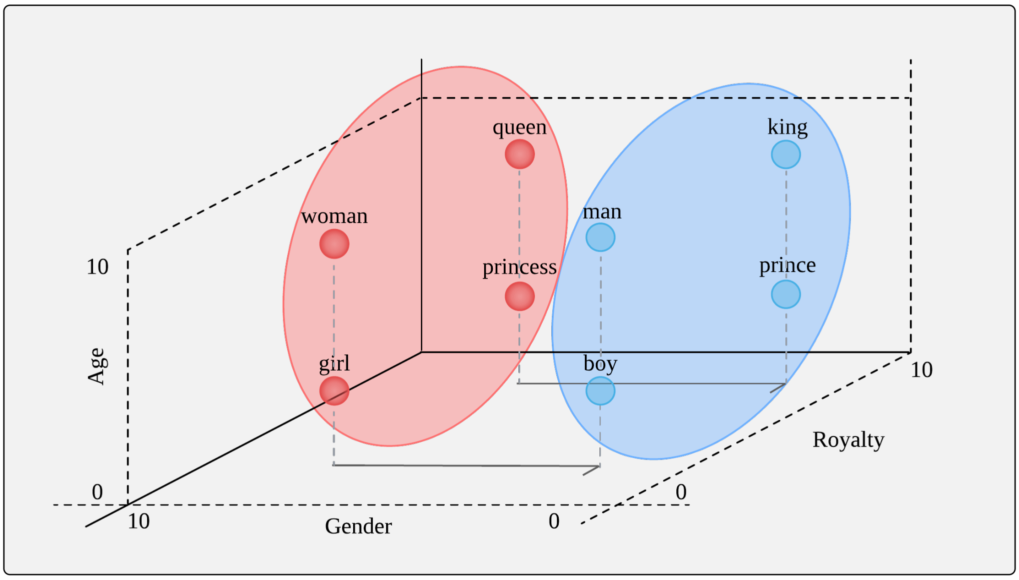

Figure 2.1: An example of a word embedding space where semantically similar words are projected by being closer in the vector space.

semantically similar short texts (determined through their word vectors) into longer, pseudo texts, using K-means clustering. LDA is then applied to the pseudo texts to assign their topics.

2.3 Recent Models

Three years later, Thompson and Mimno (2020) utilized the rise of large language models by using the *Bidirectional Encoder Representations from Transformers (BERT)* (Devlin et al. 2019) to cluster tokens based on their contextual vectors drawn from BERT. Since BERT is a bidirectional model, it considers both the left- and right-side context of a token, unlike Word2Vec, which is a context-free embedding space with a single embedding for each word. The authors show state of the art improvement over LDA models on any metric with at least one of their model variations.
Following this trend of applying or aligning transformers and language models for topic classification and modelling spawned the recent models widely used today, notably *BERTopic* (Grootendorst 2022) leveraging transformers and *Term Frequency Inverse Document Frequency (TF-IDF)*, ProdLDA (Srivastava and Sutton 2017), *Contextualized Topic Model (CTM)* (Bianchi et al. 2020) handling topic modelling in different languages, PromptTopic (H. Wang et al. 2023), Cross-Domain Topic Classification (Osnabrügge, Ash, and Morelli 2021) and Tweet Topic Classification (Antypas et al. 2022).

The commonality among these models resides in their utilization of transformers and language models, occasionally complemented by conventional methodologies like LDA, or alternatively exclusively relying on supervised learning, word embedding clustering, or prompting techniques.

2.4 Limitations and Criticism

Although the introduced models have gained significant popularity and sustained usage over time due to their commendable performance, they are nonetheless restrained by certain limitations. In addition to their evident technical challenges as outlined in the sections prior, these models also face criticism regarding their general thematic and linguistic validation, namely expressed by Shadrova (2021) and Schröter and Du (2022):

A) Topics, themes or categories are themselves not well-defined linguistic concepts, which demands an in-depth analytical process for them to be constructed correctly rather than extracting them as basic information from text.

B) The utilization of statistical word distributions and statistical means in general is deemed invalid for constructing and validating topics from a linguistic standpoint ("Topic modeling is incomplete" Shadrova (2021, pp. 8, 11)).

C) It is impossible to validate topic models, as "there is no concept in linguistics that would relate certain degrees of statistical distinctivity to certain qualitative aspects like goodness or coherence of topics" (Shadrova 2021, p. 11).

D) Given that language is dynamic and not static, the structure of documents, their content types, and the content itself are subject to continual change. As previously mentioned, one of the key challenges for a model is to effectively manage the constantly evolving temporal and worldly contexts, which may give rise to new topics, words, and environments. In order to adapt effectively, models must possess flexibility and undergo continuous training on up-to-date texts.

E) Topic models often lack sufficient flexibility in allowing users to adjust personal parameters, such as granularity (the number of topics) and their scope. As even topics themselves aren't well-defined linguistic concepts **(A)**, topic modelling demands a degree of personalization in order to be applied correctly on scoped use-cases.

These criticisms primarily target traditional topic models like LDA, yet statistical computation serves as the underlying principle for more contemporary methods such as embedding spaces and transformers as well. Nonetheless, I contend that some of these critiques are mitigated by the latest technologies, particularly those derived from supervised text classification. With VecTop, I aim to capitalize on these advancements to effectively address and further overcome some of the aforementioned criticisms.

3 VecTop

In this section, I propose a novel approach for labelling texts with topics called VecTop: Vector Database for Topic Search using Contextualized Word Embeddings.

The fundamental intuition underlying VecTop is as follows:

1) Periodically and continually scrape topic-labeled corpora from publicly available sources such as news pages.

2) From their contents, build a word embedding space stored in a vector database, alongside their corresponding topic labels.

3) Given an unlabeled text, create a representation of it within the same word embedding space.

4) Use k-nearest neighbors to determine the closest documents within the vector database and return their topic labels.

5) Assign these topics to the unlabeled text.

In the following, I will go into more detail concerning this general process and outline VecTop's architecture as shown in Figure 3.1.

3.1 Architecture

In this section, I will focus on presenting all the components of VecTop in an ideafull, not critical manner. I will talk about the limitations and challenges of the proposed architecture in section 3.5.

The foundation of VecTop lies in the abundance of corpora available on the internet, comprising written texts with assigned topics . For instance, platforms like the New York Times publish numerous human-written articles daily, organized into topics and subtopics (by the authors) to encourage user navigation and comprehension. VecTop makes use of those sources by periodically scraping the texts themselves alongside their topics and storing them.

As a next step, VecTop must represent the documents as word embeddings to capitalize on their advantages, as elucidated in Section 2.1. One approach would involve treating the entire document as a single entity and generating a single embedding vector from it. However, this method exhibits several limitations. Firstly, compressing the information of an entire document, which may consist of over 1000 words, into a single vector representation risks information loss, thereby undermining the efficacy of the topic search. Secondly, as the length of documents can vary significantly but the vectors must have fixed dimensionality,

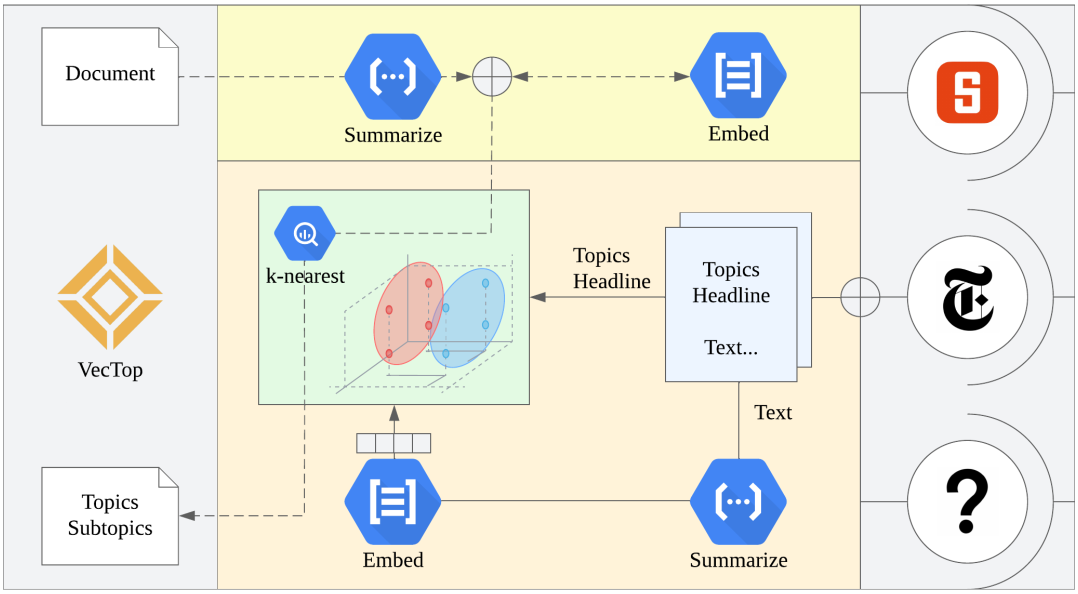

Figure 3.1: The architecture of VecTop from right to left: multiple corpora can be integrated into VecTop's ecosystem (Spiegel Online, New York Times), tailored to individual use cases. From these corpora, VecTop constructs its own vector database by initially summarizing the texts, embedding them, and then storing them alongside their corresponding topics, headlines, or breadcrumbs, depending on the corpora. When an unlabeled text requires categorization, VecTop conducts a *Topic Search* in its vector space using k-nearest neighbors, returning the topics and subtopics closest in proximity within the space, calculated by cosine similarity.

a method to truncate or pad longer and shorter documents is necessary, which again leads to potential information loss. Alternatively, a method more akin to BERTopic (Grootendorst 2022) would involve creating a single embedding for each sentence and then storing each sentence within VecTop's database. However, this approach resulted in "littering" the vector space, as it included noisy sentences devoid of contextual information and relevance to their assigned topics. This clutter made it more challenging and resource-intensive to search for sentences that are genuinely pertinent and related to their topics.

As a solution to this issue, I propose integrating TextRank (Mihalcea and Tarau 2004), an extractive text summarization technique, for the following reasons:

- Condensing the document into a fixed number of sentences and therefore a similar length, allowing VecTop to store one embedding per summary and hence document.
- Reducing noise in documents is a common challenge in topic modelling, as discussed in Section 2.1. In this context, instead of removing stopwords or focusing on keywords, TextRank identifies the most relevant sentences in their entirety based on TextRank's perspective. Keeping whole sentences is invaluable for generating contextualized word embeddings, hopefully eliminating noise while retaining context.

- In contrast to abstract text summarization methods, TextRank preserves the original wording of the text, which is desirable for this task.
- TextRank's algorithm employs a graph-voting system that, akin to LDA and other traditional methods, relies on word occurrences and frequencies. This process can be viewed as a form of preprocessing for the documents, leveraging the strengths of traditional-like methods to refine the content effectively before employing more recent technologies.

After summarizing the document, a vector embedding of 1536 dimensions is generated and stored, utilizing OpenAI's *text-embedding-ada-002*. However, alternative models such as BERT or RoBERTa (Y. Liu et al. 2019) could also be employed for this purpose.
Finally, when presented with an unlabeled text, VecTop initially summarizes its content. It then utilizes its vector database, which has been populated with various corpora and topics. A vector search is conducted using cosine similarity, finding the texts that most closely resemble the input text (k-nearest neighbors), and returning their corresponding topics and subtopics.

3.2 Corpora

Currently, VecTop offers two corpora out of the box, which have been scraped using python:

- **Spiegel Online Corpus**[1]
  A German news platform that publishes daily articles covering a wide range of topics, primarily focusing on Western news but also providing coverage of Eastern areas. The corpus currently contains more than 200.000 articles ranging from 2017 to 2023.
- **New York Times Corpus**[2]
  An American news page covering a broader range of topics than Spiegel Online, while also publishing a higher volume of articles. The corpus currently contains more than 250.000 articles ranging from 2017 to 2023.

As was alluded to, VecTop's concept can be applied to any corpus which implements the basic format outlined in Table 3.1 and the chosen corpus decides over the variety of topics and subtopics.

3.3 Results

To estimate VecTop's perfomance, I conducted a preliminary experiment using 100 speeches from deputies of the German Parliament and had VecTop determine the topics and subtopics of each speech using the *Spiegel Online Corpus* with k-nearest neighbours where $k = 5$. I then fact-checked every topic by hand. Utilizing cross-lingual word embeddings, there were

[1] https://www.spiegel.de/
[2] https://www.nytimes.com/

Table 3.1: The format of a document within a VecTop corpus.

Name	**Description**
text	The original text of the document.
summary	The TextRank summary of the *text*.
topic	The main topic this document is labeled with.
subtopic	The subtopic this document is labeled with (optional).
embedding	The embedding vector of the *summary*.
url	A reference to the source material; in case of scraped articles, this could be the original article link e.g.
date	The date this document was published on.

no discernible differences in language between the unlabeled text and the corpus during this experiment.
In the evaluation, VecTop showed a **98%** correctness on main topics and **93%** correctness on subtopics. Table 3.2 shows an excerpt of this.

3.4 Use Cases

The versatility of VecTop is contingent upon the selected corpus, rendering it applicable to various contexts. For instance, integration of VecTop into the *Bundestags-Mine* (Bönisch et al. 2023) has facilitated the categorization of all speeches delivered in the German parliament since 2017, leveraging the *Spiegel Online Corpus.* This integration enables users to effectively filter and identify speeches and representatives pertinent to their respective agendas. Another potential application involves employing VecTop by creating a corpus from scientific literature within fields such as chemistry or biology. The inherent flexibility of VecTop permits its utilization in diverse domains, provided that the prescribed corpus format is met.

3.5 Addressing the Pros and Cons

In this section, I'll assess the strengths and weaknesses of VecTop, while also discussing the criticisms mentioned in Section 2.4 as labeled from **(A)** to **(E)**.

3.5.1 Pros

1) VecTop is able to assign multiple topics and subtopics, depending on the calibration and corpus, to a single document. This allows parameterization to one's own personal use case.

2) The topics and subtopics assigned within the corpus, from which VecTop derives its results, along with the texts themselves, are generated by humans, which guarantees a certain level of quality.

Table 3.2: Exemplary outcomes regarding the assignment of topics and subtopics by VecTop. Using k-nearest neighbors, VecTop assigns multiple topics and subtopics to a single text, which has been summarized using TextRank. The final example shows an instance of falsely assigned topic classification. For showcasing purposes only, the summaries and topics have been translated to English using DeepL. Main topics are highlighted in bold, while the subtopics are designated under the respective main topic.

Summary of the unlabeled Text	**VecTop Topics**
We need efforts from companies, also supported by the works councils, we need efforts initiated by the Federal Employment Agency so that we can ensure that people are sufficiently qualified for the changed living conditions. The fact that this has been emphasised a little too much has also obscured the fact that sometimes it is "only" about decent wages. So if someone is poorly paid and has to work on today's minimum wage conditions, the minimum wage increase or a better wage overall will help them. Education and training are key and we need to create these opportunities.	**Politics** Germany **Economy** Social
Hence this comment: billions of citizens around the world have now been vaccinated, billions of people. It also protects the health of many who have pre-existing conditions, for example, who are particularly at risk in a variety of ways if they become infected. because you are confusing the citizens of this country. it is the vast majority of citizens.	**Science** Medicine
To express this in terms of guests, so that you can get a feel for what this means in this industry: in 2019, the last pre-corona year, 90 million foreign guests came to Germany; the companies with their 3 million employees now need long-term perspectives and legal flexibility in times of change; the past Christmas and New Year's Eve holidays show a completely different picture: Hotels that have far too low an occupancy rate with guests despite extensive hygiene measures, bus companies whose buses are not running and, as there are no passengers, are parked in the garage, tour operators and travel agencies that are currently feeling the reluctance of customers very strongly.	**Travel** **Economy**
Because my daughter had contact with a child who had tested positive, the health authority told me to - quote - use separate rooms within the household if possible and avoid eating meals together, and that with a two-year-old. Compulsory vaccination without sufficient safety or effectiveness of the vaccine is only one thing: clearly unconstitutional. How exactly does the vaccine behave in the body? No to compulsory vaccination!	**Education**

3) **Addressing (D)**: By leveraging continuously updated online sources like news pages, VecTop and its underlying corpus possess the ability to adapt to evolving vocabulary, language nuances, and shifting topics. This adaptability enables VecTop to remain current without the need for model retraining or significant resource allocation.

4) **Addressing (C)**: VecTop offers a level of validation by firstly using topics and subtopics from publicly available sources created by humans and secondly through a justification mechanism. As VecTop outputs its topics, it also omits the k-nearest documents within its corpus from which the topics are derived. This capability allows VecTop to justify its results by stating, "Since I've found k similar texts in my corpus, I deemed your text to be related to the following topics:".

5) **Addressing (E)**: VecTop allows a variety of parameterization, namely: the used corpus and hence topics, the determination of the level of granularity through k and the usage of a time filter ("only consider documents in the corpus since 2020").

6) VecTop requires no additional model training.

7) VecTop utilizes cross-lingual word embeddings, enhancing its performance across a diverse array of languages within the same corpus.

8) VecTop has little to no difficulties with short texts.

3.5.2 Cons

1) The topics assigned by VecTop are dependant on the used corpus, which also includes the wording and the language in general.

2) There exists an imbalance in the distribution of documents per topic, with the *Spiegel Online Corpus*, for instance, containing significantly more texts about economy than culture. This disparity may introduce bias in VecTop when it conducts topic searches. A potential solution could involve the use of paraphrasing to address this imbalance by generating paraphrased texts on topics with fewer documents, thereby augmenting the corpus and mitigating bias in VecTop's topic searches.

3) Utilizing TextRank to condense documents and standardize their sizes will most certainly result in information loss. The act of pre-processing a large document by condensing it into a few sentences must be further evaluated and potentially adjusted.

4) VecTop diverges from traditional topic modelling techniques, which limits its applicability across broader ranges of thematic fields compared to methods such as LDA. Consequently, comparing it to other technologies becomes challenging, hence the term "Topic Search" is used to characterize VecTop's functionalities.

While VecTop appears to offer more advantages than disadvantages, it is essential to acknowledge that it has not yet undergone rigorous testing in a carefully designed experiment comparing it to other topic modelling techniques. Additionally, it's worth noting that a single highlighted disadvantage may outweigh two of the showcased advantages.

4 Conclusion and Future Work

In this treatise, I introduce a novel Topic Search framework through VecTop, aiming to explore new and alternative methods of assigning topics. The discussion encompasses the historical evolution of Topic Modelling up to the present, followed by an examination of common criticisms and the subsequent presentation of VecTop.
As demonstrated in Section 3.3, the proposed framework shows promise for topic assignments, although it lacks comprehensive evaluation and testing. Notably, VecTop excels in its adaptable, lightweight environment, with access to current data and user-friendly parameterization. However, as highlighted in Section 3.5, potential errors, such as an information bottleneck introduced through TextRank or dependency on the corpora used, are identified. Despite its limitations, VecTop exhibits the potential to mitigate some of the criticisms outlined in Section 2.4 and could serve as a valuable addition to traditional Topic Modelling and Classification techniques.

Moving forward, several key steps are necessary for the advancement of VecTop. Firstly, it requires comprehensive evaluation before venturing into the exploration of additional potential corpora. Moreover, it should consider making the usage of OpenAI's *ext-embedding-ada-002* optional, while also investigating alternative word embedding models. Addressing the imbalance of topics and hence their documents could involve implementing paraphrasing options. Additionally, efforts should be directed towards rendering TextRank obsolete, perhaps by dividing documents into paragraphs of similar lengths.

Bibliography

Antypas, Dimosthenis et al. (2022). *Twitter Topic Classification.* arXiv: 2209.09824 [cs.CL].

Barde, Bhagyashree Vyankatrao and Anant Madhavrao Bainwad (2017). “An overview of topic modeling methods and tools”. In: *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 745–750. DOI: 10.1109/ICCONS.2017.8250563.

Bengio, Yoshua, Réjean Ducharme, and Pascal Vincent (2000). “A neural probabilistic language model”. In: *Advances in neural information processing systems* 13.

Bianchi, Federico, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini (2020). “Cross-lingual contextualized topic models with zero-shot learning”. In: *arXiv preprint arXiv:2004.07737*.

Blei, David and John Lafferty (2006). “Correlated topic models”. In: *Advances in neural information processing systems* 18, p. 147.

Blei, David, Andrew Ng, and Michael Jordan (2001). “Latent dirichlet allocation”. In: *Advances in neural information processing systems* 14.

Blei, David M. (Apr. 2012). “Probabilistic topic models”. In: *Commun. ACM* 55.4, pp. 77–84. ISSN: 0001-0782. DOI: 10.1145/2133806.2133826. URL: https://doi.org/10.1145/2133806.2133826.

Bönisch, Kevin, Giuseppe Abrami, Sabine Wehnert, and Alexander Mehler (2023). “Bundestags-Mine: Natural Language Processing for Extracting Key Information from Government Documents”. In: *Legal Knowledge and Information Systems.* IOS Press. ISBN: 9781643684734. DOI: 10.3233/faia230996. URL: http://dx.doi.org/10.3233/FAIA230996.

Chemudugunta, Chaitanya, Padhraic Smyth, and Mark Steyvers (2006). “Modeling general and specific aspects of documents with a probabilistic topic model”. In: *Advances in neural information processing systems* 19.

Chiang, Wei-Lin et al. (Mar. 2023). *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.* URL: https://lmsys.org/blog/2023-03-30-vicuna/.

Churchill, Rob and Lisa Singh (Nov. 2022). “The Evolution of Topic Modeling”. In: *ACM Comput. Surv.* 54.10s. ISSN: 0360-0300. DOI: 10.1145/3507900. URL: https://doi.org/10.1145/3507900.

Deerwester, Scott, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman (1990). “Indexing by latent semantic analysis”. In: *Journal of the American society for information science* 41.6, pp. 391–407.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* arXiv: 1810.04805 [cs.CL].

Ding, Chris, Tao Li, and Wei Peng (2008). “On the equivalence between Non-negative Matrix Factorization and Probabilistic Latent Semantic Indexing”. In: *Computational Statistics Data Analysis* 52.8, pp. 3913–3927. ISSN: 0167-9473. DOI: https://doi.org/10.1016/j.csda.2008.01.011. URL: https://www.sciencedirect.com/science/article/pii/S0167947308000145.

Girdhar, Yogesh, Philippe Giguère, and Gregory Dudek (2013). "Autonomous Adaptive Underwater Exploration using Online Topic Modeling". In: *Experimental Robotics: The 13th International Symposium on Experimental Robotics*. Ed. by Jaydev P. Desai, Gregory Dudek, Oussama Khatib, and Vijay Kumar. Heidelberg: Springer International Publishing, pp. 789–802. ISBN: 978-3-319-00065-7. DOI: 10.1007/978-3-319-00065-7_53. URL: https://doi.org/10.1007/978-3-319-00065-7_53.

Grootendorst, Maarten (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv: 2203.05794 [cs.CL].

Hofmann, Thomas (1999). "Probabilistic latent semantic indexing". In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57.

Kherwa, Pooja and Poonam Bansal (2019). "Topic modeling: a comprehensive review". In: *EAI Endorsed transactions on scalable information systems* 7.24.

Laureate, C.D.P., W. Buntine, and H. Linger (2023). *A systematic review of the use of topic models for short text social media analysis*. https://link.springer.com/article/10.1007/s10462-023-10471-x. [Accessed 26-02-2024].

Lee, Daniel and H Sebastian Seung (2000). "Algorithms for non-negative matrix factorization". In: *Advances in neural information processing systems* 13.

Li, Quanzhi, Sameena Shah, Xiaomo Liu, Armineh Nourbakhsh, and Rui Fang (2016). "TweetSift: Tweet Topic Classification Based on Entity Knowledge Base and Topic Enhanced Word Embedding". In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. DOI: 10.1145/2983323.2983325.

Liu, Lin, Lin Tang, Wen Dong, Shaowen Yao, and Wei Zhou (Sept. 2016). "An overview of topic modeling and its current applications in bioinformatics". In: *SpringerPlus* 5. DOI: 10.1186/s40064-016-3252-8.

Liu, Yinhan et al. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv: 1907.11692 [cs.CL].

Mihalcea, Rada and Paul Tarau (July 2004). "TextRank: Bringing Order into Text". In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Ed. by Dekang Lin and Dekai Wu. Barcelona, Spain: Association for Computational Linguistics, pp. 404–411. URL: https://aclanthology.org/W04-3252.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). *Distributed Representations of Words and Phrases and their Compositionality*. arXiv: 1310.4546 [cs.CL].

Nguyen, Dat Quoc, Richard Billingsley, Lan Du, and Mark Johnson (2015). "Improving topic models with latent feature word representations". In: *Transactions of the Association for Computational Linguistics* 3, pp. 299–313.

Nigam, Kamal, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell (2000). "Text classification from labeled and unlabeled documents using EM". In: *Machine learning* 39, pp. 103–134.

OpenAI (2022). *Introducing ChatGPT*. Accessed: 2023-06-21.

– (2023). *GPT-4 Technical Report*. arXiv: 2303.08774 [cs.CL].

Osnabrügge, Moritz, Elliott Ash, and M. Morelli (2021). "Cross-Domain Topic Classification for Political Texts". In: *Political Analysis* 31, pp. 59–80. DOI: 10.1017/pan.2021.37.

Peña, Alejandro et al. (2023). "Leveraging Large Language Models for Topic Classification in the Domain of Public Affairs". In: *Lecture Notes in Computer Science*. Springer Nature Switzerland, pp. 20–33. ISBN: 9783031414985. DOI: 10.1007/978-3-031-41498-5_2. URL: http://dx.doi.org/10.1007/978-3-031-41498-5_2.

Qiang, Jipeng, Ping Chen, Tong Wang, and Xindong Wu (2017). "Topic modeling over short texts by incorporating word embeddings". In: *Advances in Knowledge Discovery and Data Mining: 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017, Proceedings, Part II 21*. Springer, pp. 363–374.

Schröter, Julian and Keli Du (Dec. 2022). *Validating topic modeling as a method of analyzing sujet and theme*. URL: https://jcls.io/article/id/91/.

Shadrova, Anna (Oct. 2021). "Topic models do not model topics: epistemological remarks and steps towards best practices". In: *Journal of Data Mining and Digital Humanities* 2021. DOI: 10.46298/jdmdh.7595. URL: https://hal.science/hal-03261599.

Shahnaz, Farial, Michael W Berry, V Paul Pauca, and Robert J Plemmons (2006). "Document clustering using nonnegative matrix factorization". In: *Information Processing & Management* 42.2, pp. 373–386.

Sontag, David and Daniel M Roy (2009). "Complexity of inference in topic models". In: *Advances in Neural Information Processing: Workshop on Applications for Topic Models: Text and Beyond*.

Srivastava, Akash and Charles Sutton (2017). "Autoencoding variational inference for topic models". In: *arXiv preprint arXiv:1703.01488*.

Sun, Xiaofei et al. (2023). *Text Classification via Large Language Models*. arXiv: 2305.08377 [cs.CL].

Taori, Rohan et al. (2023). *Stanford Alpaca: An Instruction-following LLaMA model*.

Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal, and David M. Blei (2006). "Hierarchical Dirichlet Processes". In: *Journal of the American Statistical Association* 101.476, pp. 1566–1581. ISSN: 01621459. URL: http://www.jstor.org/stable/27639773 (visited on 02/27/2024).

Thompson, Laure and David Mimno (2020). *Topic Modeling with Contextualized Word Representation Clusters*. arXiv: 2010.12626 [cs.CL].

Vaswani, Ashish et al. (2023). *Attention Is All You Need*. arXiv: 1706.03762 [cs.CL].

Vavasis, Stephen A. (2010). "On the Complexity of Nonnegative Matrix Factorization". In: *SIAM Journal on Optimization* 20.3, pp. 1364–1377. DOI: 10.1137/070709967. eprint: https://doi.org/10.1137/070709967. URL: https://doi.org/10.1137/070709967.

Vayansky, Ike and Sathish A.P. Kumar (2020). "A review of topic modeling methods". In: *Information Systems* 94, p. 101582. ISSN: 0306-4379. DOI: https://doi.org/10.1016/j.is.2020.101582. URL: https://www.sciencedirect.com/science/article/pii/S0306437920300703.

Wang, Han et al. (2023). *Prompting Large Language Models for Topic Modeling*. arXiv: 2312.09693 [cs.AI].

Wang, Zhiqiang, Yiran Pang, and Yanbin Lin (2023). *Large Language Models Are Zero-Shot Text Classifiers*. arXiv: 2312.01044 [cs.CL].

Yan, Xiaohui, Jiafeng Guo, Shenghua Liu, Xueqi Cheng, and Yanfeng Wang (2013). "Learning topics in short texts by non-negative matrix factorization on term correlation matrix". In: *proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM, pp. 749–757.