

Martin Halleland

Topic Modeling With Text and Images Using Pretrained Embeddings

Master's thesis in Computer Science

Supervisor: Ole Jakob Mengshoel

May 2023

Martin Halleland

Topic Modeling With Text and Images Using Pretrained Embeddings

Master's thesis in Computer Science
Supervisor: Ole Jakob Mengshoel
May 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science



Abstract

The amount of data that is created and used is increasing, and finding new ways of analyzing data is becoming an increasingly more important problem. Additionally this data can appear in multiple modalities, including text and images. One research area that explores this problem is that of topic modeling.

This paper explores extending the use of pretrained embedding models to multimodal domains. We compare a new approach to traditional topic modeling methods as well as variations of the proposed method.

We show that this method is capable of finding coherent and diverse topics. These topics are comparable to other methods if slightly worse on some metrics when comparing quantitatively, but also produce interesting qualitative results.

We conclude that the approach could be useful for multimodal topic modeling, but emphasize the need for further research in metrics, including metrics that looks at more modalities, as well as exploring opportunities in a rapidly evolving field.

Sammendrag

Mengden data som blir skapt er økende, og det å finne nye måter å analysere denne dataen blir stadig et viktigere problem. Videre kan denne dataen fremkomme med flere modaliteter, som tekst og bilde. Et forskningsområde som utforsker dette problemet er det som omhandler emnemodellering.

Denne rapporten undersøker en utvidelse av bruken av forhåndstrengte representasjonsmodeller på multimodale domener. Vi sammenligner en ny tilnærming til tradisjonelle emnemodelleringsmetoder samt variasjoner av den foreslåtte metoden.

Vi viser at den nye metoden klarer å finne sammenhengende og varierte emner. Disse emnene er sammenlignbare med andre metoder, om litt dårligere på noen metrikker når man sammenligner kvantitativt, men har også noen interessante kvalitative egenskaper.

Vi konkluderer med at denne metoden kan være nyttig for multimodal emnemodellering, men vektlegger behovet for videre forskning på metrikk, spesielt med tanke på metrikker som tar hensyn til flere modaliteter, samt utforskning av nye muligheter i et fagfelt med mange nye endringer.

Preface

This document contains the master thesis for the final semester of the 5-year master's degree in Computer Science with a specialization in artificial intelligence at the Norwegian University of Science and Technology (NTNU). A preliminary study was conducted the fall of 2022, with the main work and completion of the thesis being conducted during the spring semester from January to May 2023. The supervisor for both parts of the thesis was Professor Ole Jakob Mengshoel at the Department of Computer Science.

Table of Contents

List of Figures	vi
List of Tables	vi
1 Introduction	1
1.1 Background and Motivation	1
1.2 Goals and Research Questions	2
1.3 Contributions	3
1.4 Thesis Structure	3
2 Background Theory	5
2.1 Topic Modeling	5
2.1.1 Definitions	5
2.1.2 Overview	5
2.1.3 Generative Probabilistic Models	6
2.1.4 Latent Dirichlet Allocation	6
2.1.5 Modern Natural Language Processing Models	8
2.1.6 Clustering Embeddings	10
2.2 Multimodal Models	11
2.2.1 Multimodal LDA	11
2.2.2 Text-image Models	11
2.2.3 CLIP	13
2.2.4 The Modality Gap	14

2.3	Evaluation of Topic Models	14
2.3.1	Coverage	14
2.3.2	Coherence	14
2.3.3	Qualitative	15
2.3.4	Computing Metrics	15
2.4	Additonal Related Work	16
3	Method	18
3.1	Dataset	18
3.2	Model Structure	20
3.3	Evaluation metrics	23
4	Experiments and Results	25
4.1	Experimental Plan	25
4.2	Experimental Setup	25
4.3	Experimental Results	26
5	Discussion	41
5.1	Discussion of Metrics	41
5.2	Discussion of Output	44
5.3	Discussion of Results and Research Questions	46
6	Conclusion and Future Work	48
6.1	Conclusion	48
6.2	Future Work	49

6.2.1	The WIT Dataset	49
6.2.2	Improved Metrics	49
6.2.3	Model Variations	50
	Bibliography	51

List of Figures

1	Transformer Model architecture	9
2	CLIP Training	13
3	WIT example	19
4	High Level Architecture	20
5	HDBSCAN trees for MMETopic models with chosen clusters	42
6	Projection of reduced embeddings for different modalities used in the MMETopic model	43
7	Projection of reduced embeddings for different modalities used in the MMETopic _{SBERT} variation	44

List of Tables

1	LDA performance, average across 3 runs	27
2	All topics, LDA ₅	27
3	All topics, LDA ₁₀	28
4	MMETopic performance, average across 10 runs	29
5	All topics, MMETopic ₁₀₀	30

6	All topics, MMETopic ₂₀₀	31
7	All topics, MMETopic ₄₀₀	32
8	Embedding with single modality performance, average across 10 runs	33
9	All topics, MMETopic _{TextOnly}	34
10	All topics, MMETopic _{ImageOnly}	35
11	MMETopic with alternative embedding model performance and single modality, average across 10 runs	36
12	All topics, MMETopic _{SBERT,CLIP}	37
13	All topics, MMETopic _{SBERT,TextOnly}	37
15	Topic comparison overview, two selected topics with most relevant example from each model	39
14	Performance overview with topic coherence and topic diversity, all models with best performance based on coherence where different parameters have been tested.	40

1 Introduction

This master thesis aims to explore the use of pretrained embedding models for use in multimodal topic modeling. This chapter will serve as an introduction to the rest of the report. Firstly it will go over some of the underlying background and motivation for this report's existence in Section 1.1. Then it will present the research goal and questions in Section 1.2. The main contributions will be summarized in Section 1.3. Finally it will outline the structure of the paper in Section 1.4.

1.1 Background and Motivation

The amount of data that is created every year is increasing (Hariri et al. 2019) and it becomes increasingly difficult to manage all this information. One tool that is used to explore and understand data is topic modeling. This is a type of model that tries to find underlying structures or topics in a collection of documents. Traditionally, statistical methods, such as Latent Dirichlet Allocation (LDA), have been used to find semantic probability distribution that aims to model these topics (Chauhan and Shah 2021).

In recent years, topic modeling has become more intertwined with the field of natural language processing and neural networks (Churchill and Singh 2021). In these approaches, the models aim to find distributed representations. Meaning, in the case of a neural network, that neurons participate in the representation of several topics with changes to one topic affecting the information about other topics. These representations often result in vector representations called word, sentence or document embeddings for increasing lengths of texts. Additionally these models have become increasingly more context sensitive. An older model, word2vec, would use a sliding window to learn word embeddings. More recent models build on the transformer architecture (Vaswani et al. 2017). This architecture uses an attention mechanism to use more of the sequentiality and syntax to achieve many state-of-the-art results in NLP. In topic modeling the transformers can be used for better embeddings, such as used in the BERTopic model (Grootendorst 2022).

Another problem when looking at data, is the different types of information that can appear. Often you only consider the textual information, but you can also have images, sound, metadata or sensor data. Using multiple modalities simultaneously becomes important if you want to utilize more of the available data. Perhaps one of the more common multimodal data is that of text and images. There are several approaches for working with images, but one approach that tries to also incorporate text is the CLIP model. Also spawning from the transformer architecture this model tries to align a visual and textual transformer, to achieve very interesting results (Radford et al. 2021).

In this paper, the aim is to explore these multimodal models for the use in topic modeling. The use of transformer models, especially with regards to multimodal data, is still in its early stages and there are many unanswered questions to investigate.

1.2 Goals and Research Questions

- Goal: To explore embedding-based multimodal topic models

The main goal of the thesis will be to see if you can improve topic models. More specifically I am interested in if it's possible to use pretrained embedding spaces on multimodal data for the task of topic modeling.

- Research Question 1: Is it possible to use a pretrained embedding model for multimodal topic modeling?

The first research question asks if it is possible to utilize a multimodal dataset to generate topics. Based on earlier work, it seems like it should be possible to utilize the embeddings on multimodal data to generate topics, but which challenges and results lie in these models specifically?

- Research Question 2: How does such a model as outlined in research question 1 compare to other models?

The second question asks how these topics compare to those of other models. Are they an improvement over a model that does not exploit the multimodality of the data. Do you get significantly different topics? Which new aspects can you utilize with a multimodal model that is not possible or less practical with another model. An important part of this question is also how to evaluate the different models.

1.3 Contributions

We propose a model for multimodal topic modeling that is capable of generating meaningful topics. Its performance is compared to several different models across modalities, including traditional topic modeling models and variations of the core model. It is found to perform comparably, if slightly worse, when looking at quantitative metrics, but has more impressive qualitative results.

The contributions in the paper can be summarized in the following points:

- Exploring the task of multimodal topic modeling in existing literature.
- Creating and testing a multimodal topic model able to find topics utilizing both image and text.
- Presenting and discussing various challenges and key areas for future research.

1.4 Thesis Structure

- Chapter 1 - Introduction: Explains the motivation behind this paper, including background, research goals and questions, a summarizing of the main contributions, and this overview.
- Chapter 2 - Background Theory: Contains theory used in the rest of the thesis, including materials on topic modeling, attention models, multimodal text-image models and evaluation of topic models.
- Chapter 3 - Method: Explains the method used in the paper. How the research will be conducted. Describes the overall experimental setup.

-
- Chapter 4 - Experiments and results: Explains the experiments in more detail. How the dataset has been utilized, and how the different models are used and compared.
 - Chapter 5 - Discussion: The analysis of the results and experiments, including discussion of findings and limitations.
 - Chapter 6 - Conclusion and future work: Summarizing the thesis and recommendations for future developments.

2 Background Theory

This chapter contains relevant material and theory used in this paper. Firstly, Section 2.1 will present an overview of topic modeling, what it is as well as relevant methods. Secondly, Section 2.2 will discuss multi-modality and how image and text can be related. Finally, Section 2.3 will explain how topics can be evaluated. This chapter is not meant to be an exhaustive overview, but introduces some of the most relevant or most common models or methods.

2.1 Topic Modeling

2.1.1 Definitions

Topic models M are unsupervised algorithms that takes a set of documents $D = \{d_1, d_2, \dots, d_n\}$ and outputs a set of topics $T = \{t_1, t_2, \dots, t_k\}$ in an accurate and coherent manner (Churchill and Singh 2021).

$$M(D) = T$$

The topics are often represented as a collection of words, $\{w_1, w_2, \dots, w_i\}$, that in some manner represents the related document. One can also measure how much word in the collection represents the topic and give a weight or a degree of relevance to each word.

2.1.2 Overview

There are many approaches to topic modeling. In Churchill and Singh 2021 the authors discuss the development of several topic models and different use cases. They identify four main methodologies: generative, graph-based, matrix-based and NLP-aided methods, with the most common being generative and NLP-aided. These methods can also be used in conjunction with each other.

Generative models or generative probabilistic models aim to model documents as probability distributions of topics. These are explained more in 2.1.3. NLP-aided methods aim to use approaches from other NLP tasks to improve topic modeling. Most common is to use word embeddings. This will be explained further in 2.1.5.

In addition they separate between static; one unchanging corpus, online; with the introduction of new documents and vocabulary, and temporal; with documents over time. In this paper the focus is on the use of a static corpus.

2.1.3 Generative Probabilistic Models

In generative probabilistic modeling, data is assumed to come from a generative process that includes some hidden or latent variables. If known, these latent variables would specify a probability distribution over the observed and latent variables. An example of this would be a distribution of the most likely words given a topic. Since these hidden variables cannot be known, they have to be estimated using a conditional distribution based on the observed variables. This distribution is also called the posterior distribution (Blei 2012).

The goal of generative probabilistic models is to find the posterior distribution that best fits the observed data. One such method, Latent Dirichlet Allocation, will be described in the following section.

2.1.4 Latent Dirichlet Allocation

One popular method of generative probabilistic modeling is Latent Dirichlet Allocation (LDA) introduced in Blei et al. 2003. As in other similar models, its purpose is to find the latent or hidden topics in documents from a corpus, where each topic is represented by a probability distribution of words. LDA finds these distributions by sampling a distribution and then iteratively trying to find the optimal distribution for each document.

The generative process can be described more precisely in the following fashion based on the one used in Churchill and Singh 2021 and Blei et al. 2003: Given k topics, parameters α , and β . Then for a set of documents D :

For d in D :

- (1) Randomly draw number of N words for d from the Poisson distribution
- (2) Randomly draw a topic distribution θ from the Dirchlet distribution, conditioned on α
- (3) for each w in the range of N
 - (a): Draw a topic z from a Multinomial distribution based on θ
 - (b) Draw a word w based on the probability of w given topic z , conditioned on β

One important point to notice from this generative process, is that the number of topics k is assumed known.

The optimization task becomes maximising the the log likelihood of the data given the parameters α and β . To do this you need to be able to calculate the posterior distribution, but as shown in the original paper Blei et al. 2003 this is not possible to do exactly and has to be estimated. Some common methods for this include expectation propagation (EP)(Minka 2013) and Gibbs sampling (S. Geman and D. Geman 1984)

Although LDA is a popular method, there are some key challenges with it and other generative models. First of all they are computationally expensive. To deal with this, you have to use methods of approximation. (Sontag and Roy 2012) Another problem is that low information high frequency words, commonly called stopwords, such as "the" and "and" often affect topics making them less informative and also affect metrics (see Section 2.3). Therefore when working with these methods, one must address how stopwords should be dealt with (Fan et al. 2019).

One common approach is to use various preprocessing techniques. These includes removing the stopwords before applying the method. Another is lemmatization, where different inflected forms are grouped together. One example is 'writing' and 'writes' being grouped with 'write'. Several other methods, often involving removal of special characters or certain words, also exist (Churchill and Singh n.d.).

2.1.5 Modern Natural Language Processing Models

Many newer topic models incorporate methods from areas of natural language processing (NLP). The main idea is to use prior knowledge about language to enhance the found topics. This knowledge can come in the form of a pretrained language model.

One common method is to use embeddings. Embeddings is a form of representing text in a distributed representation, for example feature vectors, such that similar vectors represent similar texts. Mikolov et al. 2013 introduced word embeddings for finding similar words with the method now called word2vec. In Le and Mikolov 2014 they extend the idea to whole documents with doc2vec.

Another recent evolution in NLP is the introduction of transformer models in Vaswani et al. 2017. This is an architecture buildings on the concept of attention. While previous efforts often represents texts as bags of words (BOW) not caring about the syntax and relationship between words, transformer models are context sensitive. This means it can understand differences between similar words used in different contexts with different meanings. The transformer model consists of a encoder, trying to encode or compress the input data, and an decoder trying to decode or extract the encoded data. Each part uses multi-head attention to figure out which part of the input information is most important for the corresponding output. These modules can be stacked and combined to create larger models. The basic architecture can be seen in Figure 1 with the encoder on the left side, and the decoder on the right.

¹<https://commons.wikimedia.org/wiki/File:The-Transformer-model-architecture.png>

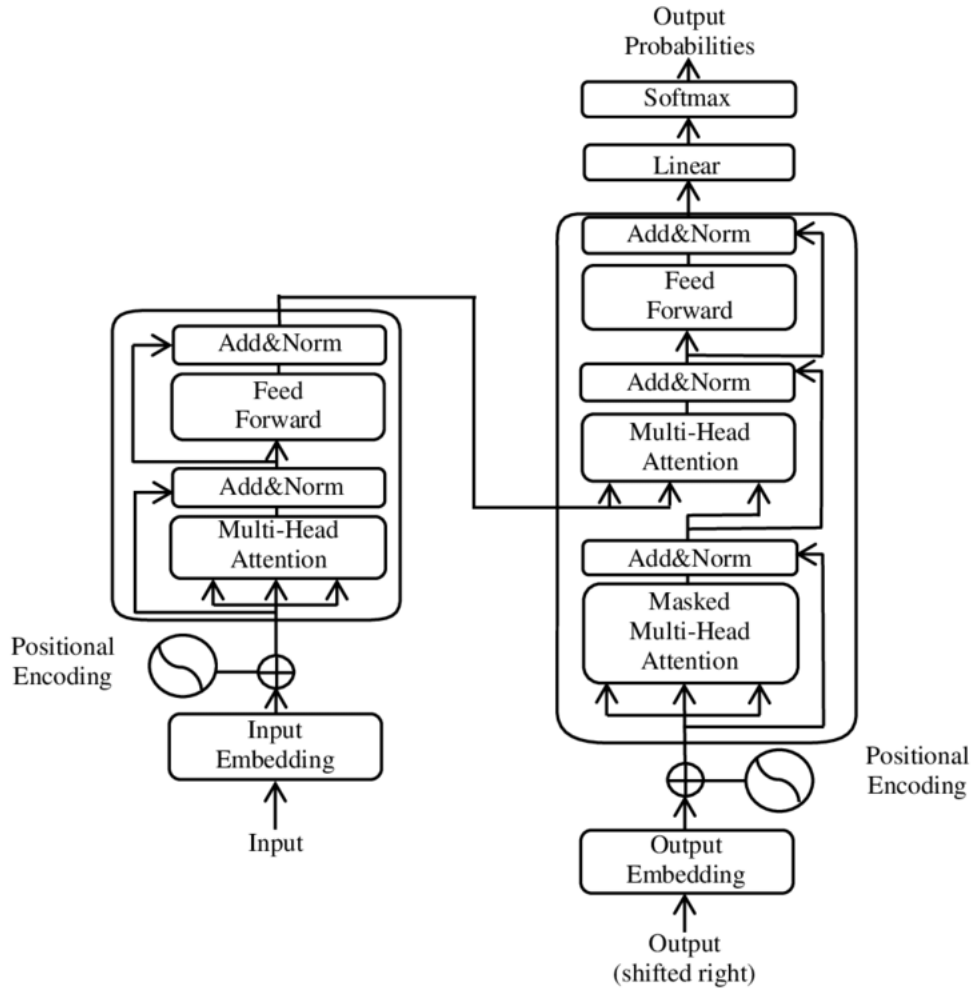


Figure 1: The transformer model architecture¹.

As a result large language models (LLMs) based on the transformer such as GPT3 Brown et al. 2020 and BERT Devlin et al. 2019 have been shown to perform well on a wide variety of NLP tasks. These models can also be used to create text embeddings.

There are different ways to utilize embeddings for topic modeling. Examples include combining them with traditional methods such as LDA, as done in Nguyen et al. 2018 and Bunk and Krestel 2018, or using clustering models such as top2vec. The latter approach will be explained in more detail in the following section, Section 2.1.6.

2.1.6 Clustering Embeddings

Top2Vec Angelov 2020 is one method that utilizes clustering on document and word embeddings to find topics. It works by first creating jointly embedded document and word vectors using one of three methods: Doc2Vec, Universal Sentence Encoder or BERTSentence Transformer. These word vectors are high-dimensional and very sparse, so to make the embedding possible to work with, you need to use methods for reducing the dimensionality. Top2Vec utilizes Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) which is a general purpose dimension reduction technique that preserves global structure from McInnes et al. 2018.

Using the lower dimensionality embedding, Top2Vec finds dense areas or clusters using hierarchical density based cluster selection (HDBSCAN) from Malzer and Baum 2020. These dense areas are categorized as the same topic. For each topic you find the centroid to find a representative called the topic vector.

From the topic vector you find the n -closest word vectors, these will then become topic words that represent the topic.

In Angelov 2020 they found that the resulting topics were more informative and representative of the used texts than probabilistic generative models.

Another method that utilizes clustering is BERTTopic (Grootendorst 2022). It is quite similar to Top2Vec also using UMAP and HDBSCAN, but is generalized to be able to use a wide variety of transformer models.

The main difference to Top2Vec is the use of the transformer embeddings. It does not use joint word and document embeddings and therefore has to find another way to represent the found topics. The way this is done is to use term-frequency inverse-document frequency (tf-idf) for each cluster or class, named class-tf-idf (c-tf-idf). Here higher scoring terms or words are chosen as representants for topics.

2.2 Multimodal Models

When working with texts or other forms of media in the real world, it is common that other forms of information are included. For example news or social media posts very often include pictures, and videos also include sound and pictures. Therefore restricting topic models to only text leaves out a huge potential source of information. Consequently there are many models that try to capture this additional multimodal information. In the subsequent sections some image-text approaches will be explained in further detail.

2.2.1 Multimodal LDA

One of the first topic models to take use of multimodal data was MoM-LDA Barnard et al. 2003. It uses a mixture of LDA models for the task of image annotation. It uses a segmentation of the image as visual words. It then assumes the visual and textual words are generated independently from the topic distribution in the LDA process.

2.2.2 Text-image Models

Another topic model that works on images and text is the Social Media based Multimodal Topic Model, SMMTM, (Huakui Zhang et al. 2022) which focuses on short social media texts with a corresponding picture. It also models text and images as bag-of-word features using feature extractors in the form of SIFT and CNNs to represent the images as visual words. The model follows a generative process similar to LDA, but with some important differences. They assume visuals can belong to multiple topics while texts only can belong to one. Another important factor they consider is if the image is consistent with the text, which is often but not always the case as the information can be complementary or contradictory. They model this in a variable called consistency. They also use a version of Gibbs Sampling in the training process.

They evaluate their model on coherence, uniqueness, and classification, using hashtags as labels, against MoM-LDA and mmETM, and outperform the other models on all metrics. They also qualitatively evaluate their models capabilities in finding topics by visualizing topics from the topic words and images and find them to be more meaningful than compared models.

2.2.3 CLIP

In Radford et al. 2021 they introduce the Contrastive Language-Image Pre-training (CLIP) model. This is a model that learns a multi-modal embeddings space by jointly training an image encoder and a text encoder to maximise the cosine similarity of the embeddings between N pairs, and minimizing it between the incorrect pairs. This training task can be seen in Figure 2 with the paired text-image pairs that are to be maximized along the diagonal. This model is able to carry out a wide variety of image related tasks with no to little additional training. These tasks include classification, image retrieval, and search.

1. Contrastive pre-training

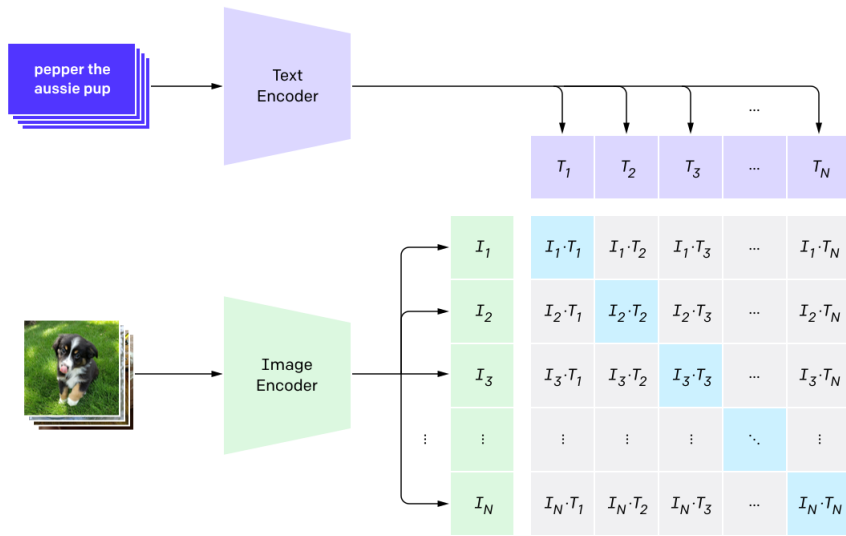


Figure 2: Contrastive language-image pretraining training process. The model is trained to find the corresponding text and images. (Radford et al. 2021)

Grootendorst 2021 has adapted the CLIP model to the task of topic modeling on images, referred to as concept modeling. This shows that the CLIP latent space is also capable of producing topics based on the images.

2.2.4 The Modality Gap

While the CLIP model tries to align text and images by maximizing cosine similarity, there is still a gap among the modalities as demonstrated in Liang et al. 2022. They show that the embeddings lie in a narrow cone, meaning even though each representation is closer to its pair than any other in the other modality, most of the same embeddings are still closer to each other. This impacts downstream performance, as well as which tasks can be performed by the model.

2.3 Evaluation of Topic Models

After using a topic model, it is important to evaluate the model to see how it performed, meaning whether or not it identified useful or good topics. Churchill and Singh 2021 identifies three general categories for evaluation of textual topics: coverage, coherence, and qualitative. These will be discussed in the following sections.

2.3.1 Coverage

Coverage is about whether or not concepts in the document collection are represented. This can be in both topics, are all topics identified, and in documents, are the assigned topics relevant to the document. Different metrics for coverage exist based on the availability of ground truth data. If such data exists, you can use metrics such as topic recall, accuracy and kl-divergence. If no such ground truth data is available, perplexity is a commonly used metric. However, perplexity and human evaluations have been shown to be poorly correlated Chang et al. 2009.

2.3.2 Coherence

Coherence aims to measure how useful individual topics are. Common metrics include precision, pointwise mutual information (PMI), Diversity, and Signal-to-Noise Ratio (SNR). Precision measures how well the most fitting topic words match the ground truth.

PMI attempts to measure closeness of words based on cofrequencies. It has become popular since it requires no ground truth (Churchill and Singh 2021) . However, in Fang et al. 2016 they find that a coherence metric based on word embeddings performed better matched with human evaluation. This measure aims to find a distance measure between the words within a topic.

Diversity looks at word overlap across all topics and sees how different these are. SNR uses ground truth to compare correct and incorrect topic words.

2.3.3 Qualitative

Qualitative evaluation methods are used to gain insight often through common sense. This can help evaluation multiple criteria at once, as well as compare human understandability of topics. Qualitative analysis has the challenge with coverage, where it is easy to choose a non representative subset of the results to be evaluated.

2.3.4 Computing Metrics

This section will go into some details on how to compute coherence using word embeddings, and how to compute diversity.

Coherence with word embeddings is calculated as an average across the topics. For each topic you look at a chosen number of words. Using all the words vector representation you find an average vector, or a centroid, representing the topic using the normalized vector representations. Then for each word in the topic you find the distance, using a metric such as cosine similarity, to the topic vector. Using these distances you find the average distance for each topic. This can be summarized in the following equation:

$$C_{we} = \frac{\sum_T \frac{\sum_{i=0}^k dist(t_c, t_{w_i})}{k}}{|T|} \quad (1)$$

Where C_{we} is the final coherence score, T the topics, k the number of chosen words from each topic, t_c the topic centroid, t_{w_i} the is a word in the topic and the *dist* function being the cosine distance.

To calculate diversity you choose a number of words for each topic to consider. Then you find the number of unique words among all topics. The diversity is the ratio of unique words to number of words considered. This can be summarized in the following equation:

$$D = \frac{|W^*|}{k * |T|} \quad (2)$$

Where D is the final diversity score, and W^* is the unique words among all of the k first words in each topic.

2.4 Additonal Related Work

In the following section some other methods working on multimodal topic modelling will be presented.

Zosa and Pivovarov 2022 propose M3L-Contrast to work on both multilingual and multimodal data. Its based on ZeroshotTM (Bianchi et al. 2021), and they utilize pretrained embeddings to find a topic distribution for each modality and language, then they train a separate network to minimize loss in Kullback-Leibler divergence between the different distributions. They evaluate topics on alignment, meaning they look for similar distributions for different languages and images, using retrieval tasks as well as evaluate topics using coherence metrics. They outperform ZeroshotTM in the multimodal setting.

Zheng et al. 2014 take a different approach to multimodal topic modeling. They build on DocNADE (Larochelle and Lauly 2012) to learn a joint representation on visual words, class labels and annotation textual words. The model they use is a neural autoregressive network. They train using a supervised method on two annotated datasets, LabelMe and UIUC-Sports. They achieve competitive performance compared to similar methods.

3 Method

This chapter describes the key components that are used in the experiments. First, the dataset that is used will be discussed, including information about the dataset as well as how it is used. Second, an overview of the multimodal topic modeling system will be given, with the core pieces of the architecture and how it is intended to work. Third, the evaluation metrics and how the models are tested will be described.

3.1 Dataset

When choosing the dataset, some key features were important. As the system that is tested works on multimodal data, the dataset of course needed to have both modalities in it. In this instance this was text and images. Furthermore, I was interested in data with text and images that were related too each other, but not necessarily a close substitution. A close substitution would for instance be an image captioning or description as in the COCO dataset(Lin et al. 2015). Another interesting feature were more substantial texts, as opposed to shorter texts like those used in social media like tweets. This would exclude datasets like Twitter100k (Hu et al. 2017) used in Huakui Zhang et al. 2022 .

One candidate that was found and subsequently used in this project was the Wikipedia-based Image Text Dataset (Srinivasan et al. 2021).

This dataset was released by Google in 2021 and is a large multimodal-multilingual dataset extracted from Wikipedia articles. It was the largest such dataset at the time at release, and is a diverse and context rich dataset. Meaning there is little repetition in a large selection of the entries in the dataset and that the texts in the dataset are descriptive, verbose and use specific terminology. They released this dataset to be able to be used in a wide variety of NLP tasks, including for use with multimodal models. Additionally, it was released along with a Kaggle competition with texts and images in an accessible format.

The full dataset contains roughly 37.1 million entries with 11.4 million unique images, and each entry having rich context in the form of image descriptions, article text, and metadata, often in multiple languages. Some of the more interesting context is the reference description which is often a description of the image, and the page description which is the Wikipedia article intro. In Figure 3 we see an example of what information can be found along with an image.

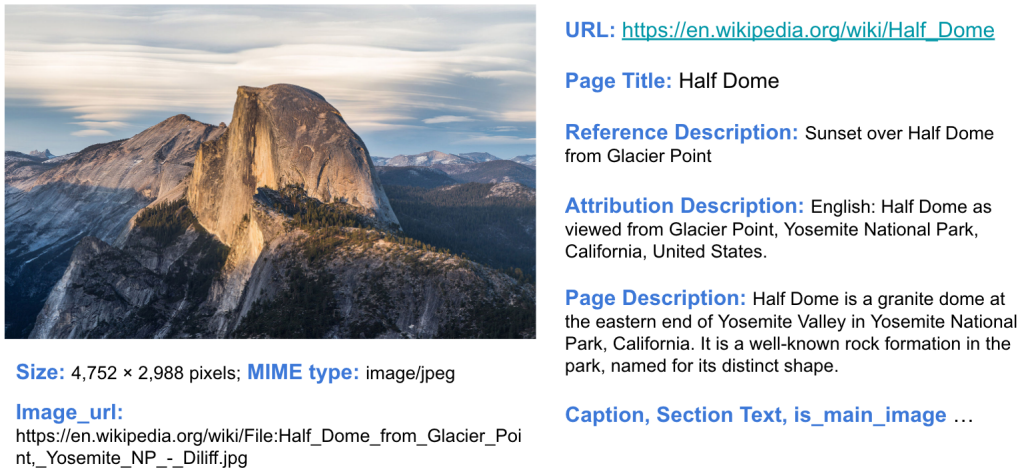


Figure 3: An example from the WIT dataset, an image along with its corresponding textual information.²

For the experiments in this paper, the full dataset is not in use. Additionally each entry is reduced to the main image of a Wikipedia article along with the page description in English. This is mostly to make the amount of data manageable to work with, as well as to focus on the most important aspect of the dataset for the system, namely the image with a related text. But more or a different part of the data could be used for a similar project.

In the end the experiments use 11548 rows of image-text entries.

²https://1.bp.blogspot.com/-JBtMJO3UGqg/YUnuRPCqW3I/AAAAAAAAIjY/QOFyL3ALtAIKYKYdPcOVfOJ8Jss_MJlowCLcBGAsYHQ/s1565/image3.png

3.2 Model Structure

The model that will be tested in the following experiments is an NLP-aided clustering based model as explained in Section 2.1.6 It builds on the idea of Grootendorst 2022 and Grootendorst 2021, but extends the model into a multimodal domain.

The model framework finds topic representations of a set of multimodal documents through four core steps. Firstly, it finds vector representations for the image and text, using an embedding model for each modality before combining them into larger representations. Then it reduces the dimensionality of the representation, before clustering the reduced vectors using a density based clustering. Using the assigned clusters, it finds topic representations using the class based variation TF-IDF. Additionally it finds image exemplars from the cluster representation to further explain found topics.

A summarization of the architecture can be seen in the high level overview in Figure 4

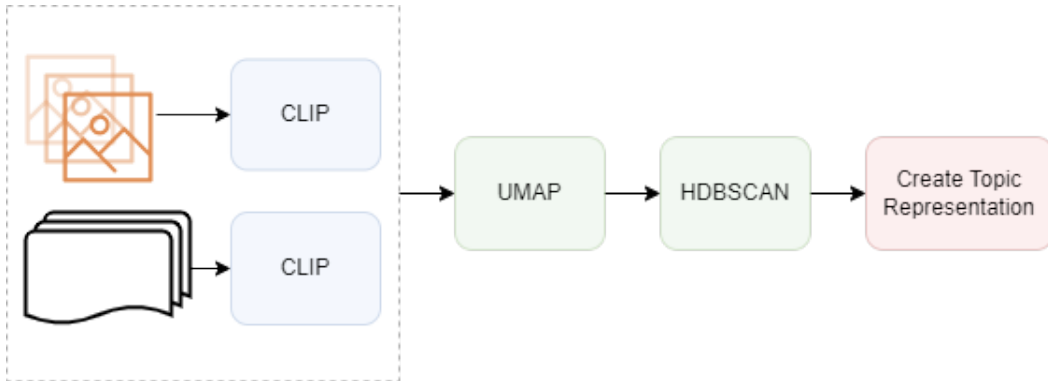


Figure 4: Architecture overview, includes creating embeddings with CLIP (Radford et al. 2021), reduce dimensionality and cluster with UMAP (McInnes et al. 2018) and HDBSCAN (Malzer and Baum 2020), and create topics from clusters

An alternative overview, with functions corresponding to the steps described in more detail in this section, can be seen in Algorithm 1.

Algorithm 1 A high abstraction level algorithm describing the proposed model

 $T, I \leftarrow \text{Texts}, \text{Images}$ $\text{textEmbedding} \leftarrow \text{EMBEDTEXT}(T)$ $\text{imageEmbedding} \leftarrow \text{EMBEDIMAGE}(I)$ $\text{combinedEmbedding} \leftarrow \text{COMBINEEMBEDDINGS}(\text{textEmbedding}, \text{imageEmbedding})$ $\text{reducedEmbedding} \leftarrow \text{REDUCEDDIMENSION}(\text{combinedEmbedding})$ $\text{clusterTree} \leftarrow \text{CLUSTEREMBEDDINGS}(\text{reducedEmbedding})$ $\text{topics} \leftarrow \text{FINDTOPICS}(\text{clusterTree}, T)$ $\text{imageExemplars} \leftarrow \text{FINDEXEMPLARS}(\text{clusterTree}, I)$

The model uses an embedding model to create vector representations of the input data, separately for the image and for the text. By using an embedding based model we assume that documents with similar topics are semantically similar and thus will have similar vector representations. The embedding model that is used can be varied, since it only affects the clustering. But in this work, transformer based architectures with variations of CLIP and BERT are used.

More specifically the models from the SentenceTransformer framework (Reimers and Gurevych 2019) are utilized. This framework provides pre-trained models with good performance with regards to embedding tasks and computation speed.

Due to the modality gap described in Section 2.2.4, one cannot cluster the image and text data as distinct data points at the same time as this would lead to meaningless clusters. Since we want to use both sources of data, we try to combine the text and image vectors by simply combining the vectors and increasing the dimensionality for each datapoint.

However it's almost impossible to work with very high dimensional data for the purposes of many clustering algorithms. Therefore it is necessary to simplify the structure with the use of a dimensionality reduction algorithm like UMAP. The goal of this algorithm is to lower the dimensionality while keeping as much of the local and global features intact as possible. UMAP achieves this in two main steps. Firstly to create a fuzzy topological representation. And then secondly to optimize a low dimensionality representation to the fuzzy topological representation measured by cross entropy. More detail about the UMAP algorithm can be found here³.

For the clustering, the HDBSCAN algorithm is used. This is an extension of the DBSCAN algorithm, with the additional features of allowing different densities in the clusters and being a hierarchical clustering algorithm. The algorithm works in five main steps. First, transform the space by finding mutual reachability distances, the max of the distance between two points and the distances of a ball around the neighbourhood of each point. Secondly, build a minimum spanning tree on these distances. Thirdly, build the cluster hierarchy by sorting the edges on distance and creating a cluster for each edge. Next, condense the tree using a minimum cluster size, denoting splits with fewer points as 'falling out' of the cluster. Finally, extract the clusters by finding the most persistent clusters when looking at the stability of the cluster splits. To define stability we introduce the inverse distance to the root λ . With λ_b being the value when the cluster was split of and became a new cluster and λ_p as the value were points fell out of a cluster. The stability for a cluster is a sum of $(\lambda_p - \lambda_b)$ for each point in the cluster. If the stability of a cluster is greater then the sum of the stabilities of its children, it is chosen as a selected cluster. This also means the number of clusters is determined by the algorithm and not by the user. More details can be be found here ⁴

³https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

⁴https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

With each datapoint assigned to a cluster, c-tf-idf is used to find topic representations for each respective cluster. C-tf-idf is based on the tf-idf or term frequency-inverse document frequency metric. The term frequency counts the number of times a term or word, t , appears in a document, d . While the inverse document is a measure of the information provided found by looking at how often the term appears across all the documents, D , and then taking the logarithm of that value. Combining these you obtain the following equation for tf-idf:

$$tfidf(t, d, D) = tf_{t,d} * \log\left(\frac{|D|}{|d \in D: t \in d|}\right)$$

This is generalized to the cluster by looking at each cluster as one large document, c , and adjusting in the following manner:

$$ctfidf(t, c) = tf_{t,c} * \log\left(1 + \frac{A}{tf_t}\right)$$

The inverse-document frequency is replaced by the inverse class frequency. Where A is the average words per class and tf_t is the frequency of a term across all classes. The 1 is added to keep the value of this term positive. Seen as a whole this value increases if a word is common in a document/class but uncommon across all documents/classes.

Additionally stopwords are ignored during this process for the same reasons as described in Section 2.1.4.

To find representative image for each cluster, we use the idea of utilizing exemplars from the HDBSCAN generated clusters from Grootendorst 2021. Here we find the most persistent points in each cluster with regards to the λ value. For each cluster these will consist of several points, so an additional selection has to be made. This is done by utilizing maximal marginal relevancy Carbonell and Goldstein 1998. This a ranking method that tries to maximize the diversity within the subset by looking at a similarity measure of the image embeddings.

3.3 Evaluation metrics

The performance of the models will be measured using two key metrics:

Coherence: measures semantic similarity within each topic

Diversity: measures semantic dissimilarity within different topics

These metrics are calculated using Equation 1 and Equation 2 from Section 2.3.4. Scoring closer to 1 on these metrics indicates a better performance. Only looking at diversity is less useful as both noise and fewer topics could lead to higher diversity, so this metric needs to be considered in relation with other evaluation methods.

Additionally qualitative assessments are performed on both found topic words and found image exemplars. This qualitative assessment will also look at coherence and diversity trying to see if they are coherent within a topic and diverse across topics.

4 Experiments and Results

In this chapter we first introduce the different experiments in Section 4.1 going over the overall experimental plan, as well as some technical details. Finally the results from the experiments are presented in Section 4.3.

4.1 Experimental Plan

The experiments were designed with the goal of studying relative performance of the proposed system, referred to as MMETopic (MultiModalEmbeddingTopic model) in the following sections, to other closely related systems including the use of different modalities. Therefore several versions of model were used on the dataset to establish several reference points for comparing the performance of the multimodal model.

Firstly a traditional LDA model was used to establish a baseline for performance on the text data. Then we ran the multimodal model, using several parameters. The best performing parameters were then used to look at the single modality performance of a similar model. Note that the textual model is essentially similar to a BERTopic model (Grootendorst 2022). Finally we also looked at the use of different embedding models for the proposed system. For each model evaluation metrics in the form of topic diversity and word embedding coherence as outlined in Section 3.3 were calculated. Additionally the output of the model in the form of topics and image exemplars were compared at qualitatively.

4.2 Experimental Setup

All experiments were performed with the dataset that is described in Section 3.1.

For the LDA model the experiments were ran with the choice of k topics $\in \{5, 10, 15\}$. For this model preprocessing was done using the OCTIS library which among other steps included removal of stopwords and lemmatization.

For the NLP-aided models the UMAP model had parameters of `n_neighbours` and `n_components` of 5, `min_dist` of 0.1 and using cosine distance for the distance metric. The HDBSCAN model were tested with `min_cluster_size` of 100, 200 and 400 for the base multimodal model and 200 for all subsequent models. Additionally, all models were run 10 times to get more robust results.

Metrics were calculated using the OCTIS library and were the ones described in Section 2.3.4.

Training was performed locally on machine with one NVIDIA GTX1060 GPU with 6GB memory and an Intel i5-7600K processor.

The code used for running the experiments, the processed dataset, as well as the results from the experiments are available on this⁵ GitHub repository.

4.3 Experimental Results

In the following section the results from the experiment are presented. There are results from the different variations in the form of tables for metrics over several runs of each model, as well as the topics found by one run that are used to qualitatively assess the different models.

In Table 1 we see the performance of the LDA model for the different choices of parameters. With Table 2 and Table 3 showing examples of topics found by the LDA₅ and LDA₁₀ versions. This model had no image part and we therefore only see the textual topic.

Table 4 shows the performance of the MMETopic model with different choices of parameters. Table 5, Table 6 and Table 7 shows the topics found for the different versions of the model.

Table 8 shows the performance for the single modality comparisons. With Table 9 showing the topics for the textual model and Table 10 showing the the topics for image model.

⁵<https://github.com/Halleland/masterDatatek>

Table 11 shows the performance of an embedding model alternative for MMETopic, as well as the performance of the text embedding only model. With Table 12 and Table 13 showing the topics for each respective model.

Table 14 shows an overview of all models for the different modalities, with the best model based on coherence were there are multiple models. Embedding based models are grouped based on the embedding model used instead of other parameters. Note that there is only one unique image-only model. Additionally, Table 15 shows an overview of two topics for each model, with one example for each topic chosen to be similar to the other models.

Model	Coherence	Diversity
LDA ₅	0.6788	0.8667
LDA ₁₀	0.7320	0.7222
LDA ₁₅	0.7064	0.7700

Table 1: LDA performance, average across 3 runs

Row	Topics LDA ₅		
0	['building', 'church', 'house', 'know', 'build', 'american', 'film', 'historic', 'band', 'locate']	['island', 'know', 'district', 'south', 'large', 'museum', 'family', 'bridge', 'river', 'year']	['station', 'play', 'railway', 'line', 'professional', 'world', 'win', 'national', 'locate', 'team']
1	['district', 'locate', 'municipality', 'county', 'town', 'north', 'population', 'south', 'region', 'area']	['united', 'states', 'county', 'census', 'population', 'city', 'township', 'state', 'place', 'war']	



Table 2: All topics, LDA₅

Row	Topics LDA ₁₀		
0	['station', 'railway', 'line', 'locate', 'building', 'build', 'church', 'house', 'historic', 'street']	['united', 'states', 'township', 'school', 'county', 'national', 'historic', 'serve', 'class', 'navy']	['film', 'know', 'american', 'band', 'album', 'singer', 'work', 'well', 'music', 'release']
1	['air', 'airport', 'force', 'year', 'saint', 'de', 'war', 'th', 'battle', 'day']	['united', 'states', 'census', 'county', 'population', 'play', 'footballer', 'city', 'area', 'professional']	['district', 'municipality', 'locate', 'town', 'population', 'village', 'city', 'area', 'south', 'region']
2	['play', 'american', 'world', 'football', 'professional', 'win', 'player', 'league', 'team', 'national']	['county', 'state', 's', 'city', 'census', 'u', 'population', 'united', 'states', 'river']	['france', 'commune', 'department', 'north', 'region', 'south', 'west', 'eastern', 'central', 'east']
3	['family', 'specie', 'genus', 'species', 'plant', 'season', 'know', 'large', 'find', 'tropical']		

Table 3: All topics, LDA₁₀




Model	Coherence	Diversity
MMETopic ₁₀₀	0.6630	0.7854
MMETopic ₂₀₀	0.6688	0.8062
MMETopic ₄₀₀	0.6306	0.8038

Table 4: MMETopic performance, average across 10 runs

Row	Topics MMETopic ₁₀₀		
0	['us' 'state' 'town' 'townships' 'census- designated' 'city' 'township' 'united' 'population' 'states' 'county' 'census'] 	['production' 'series' 'motor' 'mitsubishi' 'mercury' 'cars' 'model' 'manufac- tured' 'toyota' 'car' 'ford' 'produced'] 	['region' 'province' 'census' 'located' 'season' 'district' 'area' 'city' 'county' 'municipality' 'pop- ulation' 'tropical'] 
1	['national' 'basket- ball' 'team' 'baseball' 'former' 'player' 'league' 'professional' 'football' 'plays' 'footballer' 'played'] 	['bird' 'america' 'small' 'also' 'known' 'genus' 'species' 'nat- ive' 'common' 'plant' 'found' 'family'] 	['south' 'town' 'river' 'building' 'railway' 'county' 'located' 'built' 'station' 'his- toric' 'city' 'district'] 

2	<p>['united' 'region' 'czech' 'district' 'municipality' 'th' 'army' 'population' 'republic' 'town' 'force' 'air']</p> 	<p>['former' 'womens' 'championships' 'cup' 'player' 'football' 'national' 'league' 'team' 'played' 'world' 'professional']</p> 	<p>['actor' 'member' 'known' 'politician' 'also' 'film' 'band' 'american' 'served' 'first' 'best' 'one']</p> 
---	---	--	--

Table 5: All topics, MMETopic₁₀₀

Row	Topics MMETopic ₂₀₀		
0	<p>['tropical' 'town' 'state' 'municipality' 'township' 'area' 'city' 'population' 'county' 'states' 'united' 'census']</p> 	<p>['draft' 'played' 'baseball' 'college' 'league' 'football' 'former' 'basketball' 'professional' 'player' 'hockey' 'american']</p> 	<p>['population' 'located' 'one' 'known' 'czech' 'th' 'municipality' 'army' 'district' 'also' 'air' 'republic']</p> 

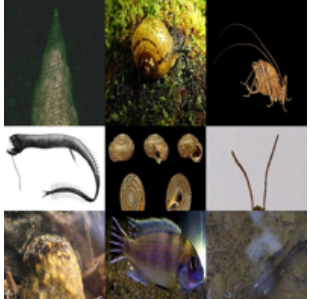




<p>1</p>	<p>['small' 'also' 'bird' 'america' 'common' 'native' 'plant' 'fam- ily' 'known' 'genus' 'species' 'found']</p> 	<p>['uss' 'first' 'world' 'class' 'aircraft' 'states' 'air' 'ship' 'war' 'built' 'navy' 'united']</p> 	<p>['church' 'area' 'loc- ated' 'city' 'river' 'dis- trict' 'building' 'his- toric' 'station' 'rail- way' 'town' 'county']</p> 
<p>2</p>	<p>['national' 'wo- mens' 'played' 'mid- fielder' 'player' 'club' 'former' 'football' 'plays' 'professional' 'footballer' 'team']</p> 	<p>['one' 'politician' 'actor' 'served' 'mem- ber' 'first' 'known' 'band' 'also' 'best' 'film' 'american']</p> 	

Table 6: All topics, MMETopic₂₀₀






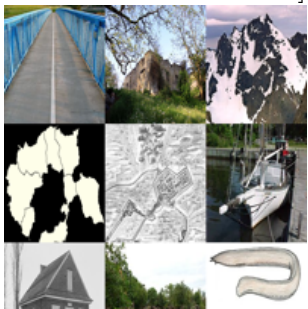

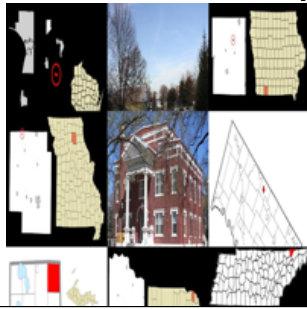



Row	Topics MMETopic ₄₀₀		
0	['small' 'bird' 'common' 'also' 'plant' 'native' 'america' 'known' 'found' 'genus' 'species' 'family'] 	['township' 'district' 'state' 'town' 'area' 'united' 'census' 'municipality' 'county' 'popula- tion' 'states' 'city'] 	['united' 'region' 'located' 'popula- tion' 'czech' 'town' 'th' 'municipal- ity' 'air' 'army' 'district' 'republic'] 
1	['south' 'building' 'railway' 'district' 'built' 'city' 'historic' 'located' 'county' 'river' 'town' 'station'] 	['one' 'played' 'former' 'member' 'band' 'film' 'american' 'first' 'best' 'known' 'world' 'also'] 	

Table 7: All topics, MMETopic₄₀₀




Model	Coherence	Diversity
MMETopic _{TextOnly}	0.7080	0.8364
MMETopic _{ImageOnly}	0.6783	0.8182

Table 8: Embedding with single modality performance, average across 10 runs

Row	Topics MMETopic _{TextOnly}		
0	['area' 'new' 'county' 'built' 'city' 'first' 'known' 'loc- ated' 'one' 'also'] 	['world' 'car' 'ship' 'aircraft' 'uss' 'war' 'designed' 'pro- duced' 'built' 'navy'] 	['historic' 'town' 'city' 'located' 'census' 'county' 'population' 'states' 'united' 'township'] 
1	['common' 'amer- ica' 'known' 'native' 'bird' 'endemic' 'found' 'genus' 'family' 'species'] 	['eastern' 'northcent- ral' 'northeastern' 'northern' 'northwest- ern' 'southwestern' 'region' 'department' 'commune' 'france'] 	['city' 'germany' 'km' 'town' 'province' 'village' 'region' 'municipality' 'loc- ated' 'district'] 

2	['basketball' 'former' 'football' 'team' 'player' 'league' 'plays' 'footballer' 'played' 'professional']	['member' 'french' 'singer' 'known' 'american' 'actor' 'actress' 'band' 'politician' 'best']	
			

Table 9: All topics, $\text{MMETopic}_{\text{TextOnly}}$

Row	Topics $\text{MMETopic}_{\text{ImageOnly}}$		
0	['building' 'river' 'town' 'historic' 'railway' 'city' 'county' 'located' 'station' 'district']	['first' 'world' 'air- craft' 'air' 'united' 'states' 'war' 'built' 'navy' 'ship']	['town' 'municipal- ity' 'area' 'township' 'city' 'united' 'pop- ulation' 'census' 'states' 'county']
			









<p>1</p>	<p>['south' 'united' 'one' 'family' 'known' 'also' 'municipality' 'genus' 'species' 'found']</p> 	<p>['national' 'world' 'played' 'football' 'basketball' 'team' 'professional' 'player' 'plays' 'footballer']</p> 	<p>['politician' 'member' 'served' 'best' 'amer- ican' 'band' 'also' 'film' 'first' 'known']</p> 
----------	--	---	--

Table 10: All topics, $MME_{Topic_{ImageOnly}}$

Model	Coherence	Diversity
MMETopic _{SBERT}	0.6832	0.8214
MMETopic _{TextOnly,SBERT}	0.6371	0.9323

Table 11: MMETopic with alternative embedding model performance and single modality, average across 10 runs

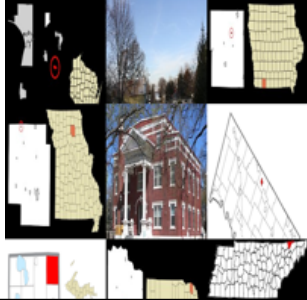



Row	Topics MMETopic _{SBERT,CLIP}		
0	['district' 'state' 'town' 'municipal- ity' 'county' 'census' 'city' 'township' 'united' 'states' 'area' 'population'] 	['force' 'uss' 'first' 'air' 'aircraft' 'world' 'states' 'ship' 'built' 'united' 'navy' 'war'] 	['area' 'south' 'dis- trict' 'railway' 'county' 'station' 'historic' 'loc- ated' 'building' 'city' 'town' 'river'] 
1	['football' 'world' 'hockey' 'basketball' 'national' 'league' 'played' 'footballer' 'player' 'profes- sional' 'team' 'plays'] 	['one' 'actor' 'politi- cian' 'member' 'band' 'served' 'best' 'first' 'film' 'also' 'known' 'american'] 	['region' 'located' 'czech' 'town' 'th' 'force' 'repub- lic' 'population' 'air' 'army' 'dis- trict' 'municipality'] 

2	<p>['also' 'common' 'de- scribed' 'america' 'extinct' 'genus' 'fish' 'family' 'species' 'known' 'found' 'sea']</p> 	
---	--	--

Table 12: All topics, $\text{MMETopic}_{\text{SBERT},\text{CLIP}}$

Row	Topics $\text{MMETopic}_{\text{SBERT},\text{TextOnly}}$		
0	<p>['endemic' 'small' 'found' 'america' 'genus' 'also' 'plant' 'species' 'native' 'com- mon' 'family' 'known']</p> 	<p>['one' 'former' 'war' 'played' 'film' 'united' 'world' 'first' 'pro- fessional' 'also' 'american' 'known']</p> 	<p>['united' 'states' 'area' 'municipal- ity' 'town' 'census' 'county' 'station' 'district' 'located' 'population' 'city']</p> 

Table 13: All topics, $\text{MMETopic}_{\text{SBERT},\text{TextOnly}}$

Topic	LDA	MMETopic _{TextOnly}	MMETopic _{ImageOnly}
0	['united', 'states', 'census', 'county', 'population', 'play', 'footballer', 'city', 'area', 'professional']	['historic' 'town' 'city' 'located' 'census' 'county' 'population' 'states' 'united' 'township'] 	['town' 'municipal-ity' 'area' 'township' 'city' 'united' 'pop-ulation' 'census' 'states' 'county'] 
1	['family', 'specie', 'genus', 'species', 'plant', 'season', 'know', 'large', 'find', 'tropical']	['common' 'amer-ica' 'known' 'native' 'bird' 'endemic' 'found' 'genus' 'family' 'species'] 	['south' 'united' 'one' 'family' 'known' 'also' 'municipality' 'genus' 'species' 'found'] 




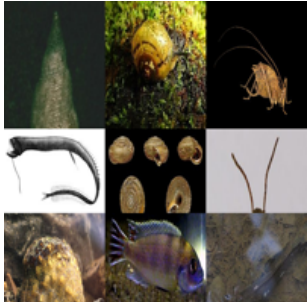

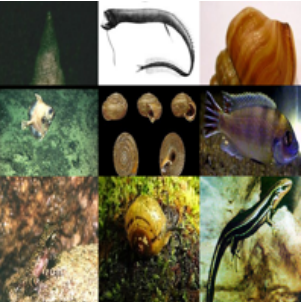
Topic	MMETopic _{CLIP}	MMETopic _{SBERT,Text}	MMETopic _{SBERT,CLIP}
0	['tropical' 'town' 'state' 'municipal- ity' 'township' 'area' 'city' 'population' 'county' 'states' 'united' 'census'] 	['united' 'states' 'area' 'municipal- ity' 'town' 'census' 'county' 'station' 'district' 'located' 'population' 'city'] 	['district' 'state' 'town' 'municipal- ity' 'county' 'census' 'city' 'township' 'united' 'states' 'area' 'population'] 
1	['small' 'also' 'bird' 'america' 'common' 'native' 'plant' 'fam- ily' 'known' 'genus' 'species' 'found'] 	['endemic' 'small' 'found' 'america' 'genus' 'also' 'plant' 'species' 'native' 'com- mon' 'family' 'known'] 	['also' 'common' 'de- scribed' 'america' 'extinct' 'genus' 'fish' 'family' 'species' 'known' 'found' 'sea'] 

Table 15: Topic comparison overview, two selected topics with most relevant example from each model

	Text-only		Image-only		Multimodal	
	TC	TD	TC	TD	TC	TD
LDA	0.7320	0.7222	-	-	-	-
MMETopic _{CLIP}	0.7080	0.8364	0.6783	0.8182	0.6688	0.8062
MMETopic _{SBERT,CLIP}	0.6371	0.9323	0.6783	0.8182	0.6832	0.8214

Table 14: Performance overview with topic coherence and topic diversity, all models with best performance based on coherence where different parameters have been tested.

5 Discussion

In the chapter the results from the previous chapter will be analyzed and discussed in more detail. First, the quantitative metrics will be discussed in Section 5.1. Then the output will be looked at for a qualitative evaluation in Section 5.2. Finally, in Section 5.3 the results will be discussed in relation to the research questions from Section 1.2.

5.1 Discussion of Metrics

Looking at the overall metrics we see that the overall most coherent model is LDA₁₀. We see that both more and less topics, affect the coherence. This might be because with fewer topics, one topic might try to incorporate too much, we can see hints of this in Table 2. Having too many topics does not have this issue, but we might instead introduce overlapping topics, such as the first and third topics in row 1 in Table 5. This obviously affects diversity, but might also affect coherence, as the differences between similar topics can come from uncommon combinations which affects coherence.

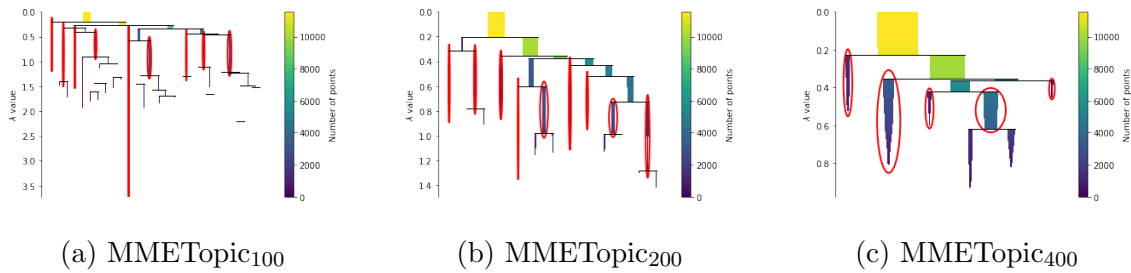


Figure 5: HDBSCAN trees for MMETopic models with chosen clusters

We also see that all the MMETopics versions are quite close to each other. With minimum cluster size of 200 we get around 10 topics, (8 for the example run). With 100 we get 9 topics and a very similar score. With 400 we see the score falling quite a bit, and it only has 5 topics. We can also look at the hierarchical tree created by HDBSCAN algorithm. This shows which clusters were chosen, their size, and where they diverged from other clusters. Figure 5a and Figure 5b shows several quite balanced clusters with regards to their size, with Figure 5b having quite a few candidates that were not chosen. These smaller branches increases the complexity of the tree, but might be a result of variance. This suggests a higher minimum size for the clusters is appropriate. Figure 5c has fewer clusters of more varied size. The impact on the coherence score might suggest that these larger clusters resulted in worse topics.

When comparing the metrics to the single modality models, Table 8, we see that both of them are slightly higher in both coherence and diversity. The text-only being higher is more expected since the metrics focuses on the text, while in the multimodal model we try to also use the visual information to create topics. But the high score for the image-only is less expected. This might mean the image clusters were quite similar to the text-only clusters. This could be because the dataset was chosen to have similar info in text and images and that this held for similar texts, meaning similar texts also had similar images. You could also look at larger or different part of the dataset, to see if this performance can be replicated there.

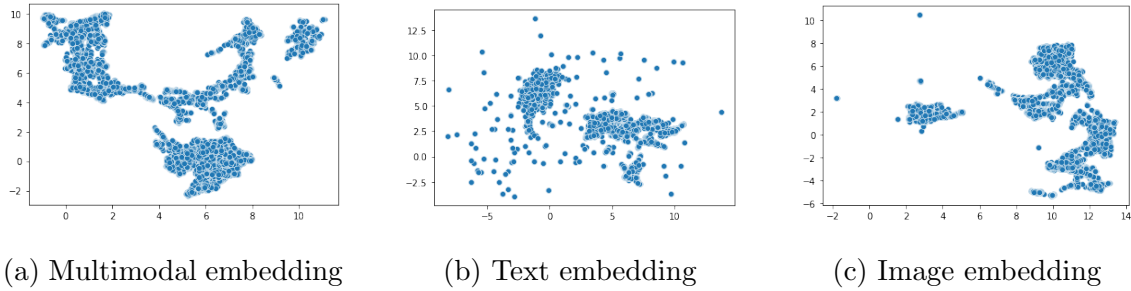
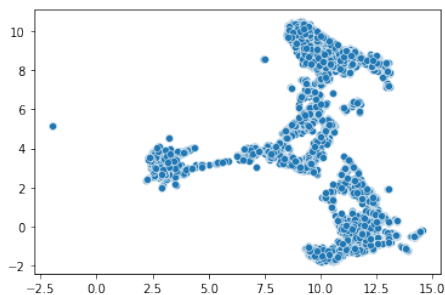


Figure 6: Projection of reduced embeddings for different modalities used in the MMETopic model

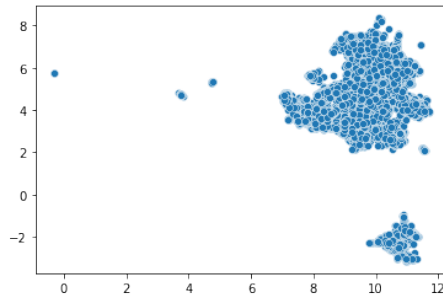
We can also look closer and compare projections of the reduced embeddings of the MMETopic model to the single modalities. In Figure 6a one can see that the clusters are more spread out while in Figure 6b and Figure 6c we get more distinct clusters, even though Figure 6b appears to be more noisy. These distinct clusters might explain the higher coherence scores as that means both single modality models found groups with similar characteristics.

The models scoring highest on diversity are the LDA_5 and $MMETopic_{TextOnly,SBERT}$ models. As mentioned fewer topics can often lead to increased diversity since there is less chance of overlap, and these had 5 and 3 topics respectively. Excluding these, we see that most of the clustering models score quite similarly, while the LDA variations are quite a bit behind. It makes sense that these models have higher diversity as to be considered for a topic, there needs to be a large distance between clusters and words that are unique for each cluster are considered to a larger degree.

It is interesting to see that the $MMETopic_{SBERT}$ outperforms the MMETopic on both metrics even with the text only version being considerably worse. The reason for the performance of the text only might be the more dense clusters as seen in Figure 7b, meaning the chosen parameters weren't able to find distinct enough clusters. For the multimodal version it's hard to say why it outperforms. Often you see that an average of multiple models outperforms a single model Wang et al. 2022. Using two different base models, might have had the same effect here.



(a) Multimodal SBERT embedding



(b) Text SBERT embedding

Figure 7: Projection of reduced embeddings for different modalities used in the $\text{MMETopic}_{\text{SBERT}}$ variation

5.2 Discussion of Output

When looking at the output we can look at both the chosen topic words and the image exemplars for the different models. Starting with the textual output of the LDA model in Table 3, it becomes clear why this model had lower diversity score. Several of the topics show signs of overlapping topics such as geographical and sociological information about areas of the United States.

Looking at the output of MMETopic models we see that the models are able to find very distinct topics with regards to both images and texts. There is some overlap for example with geographical topics and sports overlapping somewhat with the LDA model. But when we look at the images, for example the geographical topics, they are more distinct from each other, such as the first and third topic in row 0 in Table 5. It's important to remember that the images are chosen from exemplars, meaning among the core points in a cluster, and there might be more actual overlap in each cluster where the model is not as certain.

We also see a car topic appearing only in MMETopic_{100} , in the second topic of row 0 of Table 5. This looks like quite a clear topic with all words and images being seeming related. The reason that it only appears here, might be that this is a smaller topic that gets overshadowed by other topics and does not get to be a part of the core for another topic as to not show up in other exemplars.

Some of the less clear clusters are those related to people, such as the third topic in row 2 for Table 5 and the second topic in row 1 for Table 12. Here the topic words covers quite a few different areas such as actors, politicians, and musicians, with the images often being portraits. This might be a topic covering biographies or similar entries, which is a reasonable grouping, but makes the topic less clear.

When we look at the single modality outputs we see that they perform well with regards to their respective modality. The text modality, Table 9, has varied topics, meaning there is little overlap, again a little with regards to geography. We see alot of the same topics found including navy, sports, biographies, nature and buildings. But if we look at the images they are much less coherent. For the text model these are the exemplars based only on the text clusters, so there of course no reason to see a lot of similar images apart from the fact that they should come from similar texts.

For the image model, Table 10, we again see lots of the same topics, and here the text seems as coherent as the images compared to the other models for the similar categories. This could be as mentioned because the images and texts are supposed to be linked so they likely find lots of the same clusters. However we see one topic that makes much less sense then the others that have been discussed, namely the first topic in row 1. Here the images have some similar traits, mainly the background, but otherwise are not very coherent. We see this in the textual topic as well with no obvious topic presenting itself. This could be a case were a pure image model is not ideal.

The variation MMETopic_{SBERT} is also able to find many of the topics as the other models. All the topics and related images seems reasonable, meaning there are no obvious failure cases. Additionally, in the "nature" topic we see a more specific topic with some aquatic themes showing in both words and images. For the text only version we don't see as clear results, but as discussed this is most likely due to the number of topics.

Overall we see that the multimodal models are able to find similar topics to the single modality models, but are able to additionally extract image exemplars that are more coherent while being similar to the textual topic. One challenge that has not been looked at is how well the images in the core of the cluster represent the cluster as a whole. This would require new metrics or methods.

5.3 Discussion of Results and Research Questions

- Research Question 1: Is it possible to use a pretrained embedding model for multimodal topic modeling?

Based on the results from the experiments in Section 4 and the literature review it seems very possible to use pretrained embedding models for multimodal topic modeling. Looking at Table 4 and figures Table 5, Table 6 and Table 7 we see that we get meaningful output from the model with word topics and image exemplars being both coherent and diverse as described in Section 3.3.

However, there is a challenge in the modality gap described in Liang et al. 2022. Here we used a direct approach of combing the embeddings to work around this problem, but other methods could be used. One will be discussed in the future work section, Section 6.2.

- Research Question 2: How does such a model as outlined in research question 1 compare to other models?

Though we conclude for RQ1 that it is possible to use the proposed model to take a multimodal approach to topic modeling, this does not mean it is necessarily better. From the previous discussion we see that using single modalities, both for the traditional and embedding based methods, score higher on coherence. The multimodal method is not far off, however, and additionally is able to identify coherent image clusters for the topics. We also see that even though the image only model is able to find similar image clusters, it looks more prone to worse topics. Therefore the proposed system could be useful in cases were images are an important factor in the topic generation.

In addition we find that using two different embedding models for the image and textual mode outperformed using CLIP for both modalities. This performance was quite close both in metrics and output, so it is difficult to say where this difference stems from. In Section 5.1 we propose that this might be an averaging effect.

6 Conclusion and Future Work

This chapter is gonna provide a conclusion of the thesis and present some future work opportunities for multimodal topic modeling.

6.1 Conclusion

Methods of communication are rapidly changing and often require more and more data. We are not only limited to text, but use images, sound and video. Therefore it is not a good idea to limit the modalities that are used in fields like topic modeling. This thesis has studied one way of using text and images for multimodal topic modeling with the use of embedding models. The WIT dataset was utilized due to the context rich images-text pairs which made it a useful dataset for topic modeling. Building on the ideas in BERTTopic, the MMETopic model was created with a couple of variations. The performance of this model, and related models using single modalities, and an LDA model was tested and compared using quantitative metrics as well as qualitative analysis of the model outputs. It was shown that it was possible to use the proposed system to find suitable topics, but with worse performance than the LDA baseline when looking at coherence, and slightly worse than the single modality embedding models when looking a both coherence and diversity. However, the multimodal models were able to find useful information in the images and could present image clusters that looked representative of the topics, and to a larger degree than the single modality embedding models. That means this type of model could be useful in applications were using both images and text is important. These images were only evaluated qualitatively and a need for better metrics regarding images in topic modeling was identified.

6.2 Future Work

6.2.1 The WIT Dataset

The WIT dataset that was used has a lot of opportunities to be explored further. In this thesis we only used a small part of the full dataset, as well as just the most important parts for each entry. It could be possible to use the additional information that exists in some interesting way. Of course you could also use different parts of the dataset to see if similar results as found in this thesis holds, and use a larger part to test scalability of different systems.

6.2.2 Improved Metrics

As mentioned earlier in the thesis there is a need for better metrics for topic modeling, especially in a multimodal setting. Coherence and diversity metrics for images calculated in a similar manner as for the texts could probably be used, but these would have to be tested up against human evaluation as it's not always easy to see which metrics correlate well as demonstrated by the perplexity metric. More importantly for multimodal models is some metrics that can describe the visual information, and its connection to the text. For example how well chosen visual info and textual topic words correspond to each other.

Another related point is how to use the visual information. Here we used image exemplars, but in the related work we saw others using visual words. Yet another method could be to somehow utilize graph methods as in Yang et al. 2018 to find common elements in a different manner.

6.2.3 Model Variations

Liang et al. 2022 demonstrate a modality gap in the CLIP embeddings. However, So et al. 2022 try to remove this gap by using a Mixup (Hongyi Zhang et al. 2018) data augmentation and introducing interpolated hard negative samples to decrease the gap between pairs and increase robustness. Similar methods could make it possible to treat text and images as separate data points and cluster them as one larger dataset. How you categorize topics becomes an even more important part for this method.

As seen with OpenAI 2023 multimodal models are rapidly changing and new opportunities and approaches for topic modeling could present themselves. Several models trying to connect text and images, such as Midjourney⁶, DALL-E⁷ and Stable Diffusion⁸, many building on CLIP, could probably be utilized in different manners to improve topic models in a similar fashion as attempted in this thesis. It is therefore important to keep up to date in the field.

⁶<https://www.midjourney.com/home/>

⁷<https://openai.com/product/dall-e-2>

⁸<https://stability.ai/stable-diffusion>

Bibliography

- Angelov, Dimo (2020). *Top2Vec: Distributed Representations of Topics*. DOI: 10.48550/ARXIV.2008.09470. URL: <https://arxiv.org/abs/2008.09470>.
- Barnard, Kobus et al. (2003). ‘Matching words and pictures’. In: *The Journal of Machine Learning Research* 3, pp. 1107–1135.
- Bianchi, Federico et al. (Apr. 2021). ‘Cross-lingual Contextualized Topic Models with Zero-shot Learning’. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 1676–1683. DOI: 10.18653/v1/2021.eacl-main.143. URL: <https://aclanthology.org/2021.eacl-main.143>.
- 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>.
- Blei, David M., Andrew Y. Ng and Michael I. Jordan (Mar. 2003). ‘Latent Dirichlet Allocation’. In: *J. Mach. Learn. Res.* 3.null, pp. 993–1022. ISSN: 1532-4435.
- Brown, Tom B. et al. (2020). *Language Models are Few-Shot Learners*. DOI: 10.48550/ARXIV.2005.14165. URL: <https://arxiv.org/abs/2005.14165>.
- Bunk, Stefan and Ralf Krestel (2018). ‘WELDA: Enhancing Topic Models by Incorporating Local Word Context’. In: *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries. JCDL ’18*. Fort Worth, Texas, USA: Association for Computing Machinery, pp. 293–302. ISBN: 9781450351782. DOI: 10.1145/3197026.3197043. URL: <https://doi.org/10.1145/3197026.3197043>.
- Carbonell, Jaime and Jade Goldstein (1998). ‘The use of MMR, diversity-based reranking for reordering documents and producing summaries’. In: *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 335–336.
- Chang, Jonathan et al. (2009). ‘Reading Tea Leaves: How Humans Interpret Topic Models’. In: *Advances in Neural Information Processing Systems*. Ed. by Y. Bengio et al. Vol. 22. Curran Associates, Inc. URL: <https://proceedings.neurips.cc/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf>.

-
- Chauhan, Uttam and Apurva Shah (Sept. 2021). ‘Topic Modeling Using Latent Dirichlet Allocation: A Survey’. In: *ACM Comput. Surv.* 54.7. ISSN: 0360-0300. DOI: 10.1145/3462478. URL: <https://doi.org/10.1145/3462478>.
- Churchill, Rob and Lisa Singh (Dec. 2021). ‘The Evolution of Topic Modeling’. In: *ACM Comput. Surv.* Just Accepted. ISSN: 0360-0300. DOI: 10.1145/3507900. URL: <https://doi.org/10.1145/3507900>.
- (n.d.). ‘textPrep: A Text Preprocessing Toolkit for Topic Modeling on Social Media Data [textPrep: A Text Preprocessing Toolkit for Topic Modeling on Social Media Data]’. In: *Proceedings of the 10th International Conference on Data Science, Technology and Applications* (). DOI: 10.5220/0010559000600070. URL: <https://par.nsf.gov/biblio/10280456>.
- Devlin, Jacob et al. (June 2019). ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: <https://aclanthology.org/N19-1423>.
- Fan, Angela, Finale Doshi velez and Luke Miratrix (May 2019). ‘Assessing topic model relevance: Evaluation and informative priors’. In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 12. DOI: 10.1002/sam.11415.
- Fang, Anjie et al. (2016). ‘Using Word Embedding to Evaluate the Coherence of Topics from Twitter Data’. In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’16. Pisa, Italy: Association for Computing Machinery, pp. 1057–1060. ISBN: 9781450340694. DOI: 10.1145/2911451.2914729. URL: <https://doi.org/10.1145/2911451.2914729>.
- Geman, Stuart and Donald Geman (1984). ‘Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images’. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6.6, pp. 721–741. DOI: 10.1109/TPAMI.1984.4767596.
- Grootendorst, Maarten (2021). *MaartenGr/Concept: Concept Modeling: Topic Modeling on Images and Text*. <https://github.com/MaartenGr/Concept>. (Accessed on 11/2022).
-

-
- Grootendorst, Maarten (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. DOI: 10.48550/ARXIV.2203.05794. URL: <https://arxiv.org/abs/2203.05794>.
- Hariri, Reihaneh H, Erik M Fredericks and Kate M Bowers (2019). ‘Uncertainty in big data analytics: survey, opportunities, and challenges’. In: *Journal of Big Data* 6.1, pp. 1–16.
- Hu, Yuting et al. (2017). *Twitter100k: A Real-world Dataset for Weakly Supervised Cross-Media Retrieval*. DOI: 10.48550/ARXIV.1703.06618. URL: <https://arxiv.org/abs/1703.06618>.
- Larochelle, Hugo and Stanislas Lauly (2012). ‘A Neural Autoregressive Topic Model’. In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2012/file/b495ce63ede0f4efc9eec62cb947c162-Paper.pdf.
- Le, Quoc V. and Tomas Mikolov (2014). *Distributed Representations of Sentences and Documents*. DOI: 10.48550/ARXIV.1405.4053. URL: <https://arxiv.org/abs/1405.4053>.
- Liang, Weixin et al. (2022). *Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning*. arXiv: 2203.02053 [cs.CL].
- Lin, Tsung-Yi et al. (2015). *Microsoft COCO: Common Objects in Context*. arXiv: 1405.0312 [cs.CV].
- Malzer, Claudia and Marcus Baum (Sept. 2020). ‘A Hybrid Approach To Hierarchical Density-based Cluster Selection’. In: *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE. DOI: 10.1109/mfi49285.2020.9235263. URL: <https://doi.org/10.1109%2Fmfi49285.2020.9235263>.
- McInnes, Leland, John Healy and James Melville (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. DOI: 10.48550/ARXIV.1802.03426. URL: <https://arxiv.org/abs/1802.03426>.
- Mikolov, Tomas et al. (2013). *Efficient Estimation of Word Representations in Vector Space*. DOI: 10.48550/ARXIV.1301.3781. URL: <https://arxiv.org/abs/1301.3781>.

-
- Minka, Thomas P. (2013). *Expectation Propagation for approximate Bayesian inference*. DOI: 10.48550/ARXIV.1301.2294. URL: <https://arxiv.org/abs/1301.2294>.
- Nguyen, Dat Quoc et al. (2018). ‘Improving Topic Models with Latent Feature Word Representations’. In: DOI: 10.48550/ARXIV.1810.06306. URL: <https://arxiv.org/abs/1810.06306>.
- OpenAI (2023). *GPT-4 Technical Report*. arXiv: 2303.08774 [cs.CL].
- Radford, Alec et al. (2021). *Learning Transferable Visual Models From Natural Language Supervision*. DOI: 10.48550/ARXIV.2103.00020. URL: <https://arxiv.org/abs/2103.00020>.
- Reimers, Nils and Iryna Gurevych (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. arXiv: 1908.10084 [cs.CL].
- So, Junhyuk et al. (2022). *Geodesic Multi-Modal Mixup for Robust Fine-Tuning*. arXiv: 2203.03897 [cs.CV].
- Sontag, David and Daniel Roy (May 2012). ‘Complexity of Inference in Latent Dirichlet Allocation’. In.
- Srinivasan, Krishna et al. (2021). ‘WIT: Wikipedia-Based Image Text Dataset for Multimodal Multilingual Machine Learning’. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’21. Virtual Event, Canada: Association for Computing Machinery, pp. 2443–2449. ISBN: 9781450380379. DOI: 10.1145/3404835.3463257. URL: <https://doi.org/10.1145/3404835.3463257>.
- Vaswani, Ashish et al. (2017). *Attention Is All You Need*. DOI: 10.48550/ARXIV.1706.03762. URL: <https://arxiv.org/abs/1706.03762>.
- Wang, Xiaofang et al. (2022). *Wisdom of Committees: An Overlooked Approach To Faster and More Accurate Models*. arXiv: 2012.01988 [cs.CV].
- Yang, Xu et al. (2018). *Auto-Encoding Scene Graphs for Image Captioning*. arXiv: 1812.02378 [cs.CV].
- Zhang, Hongyi et al. (2018). *mixup: Beyond Empirical Risk Minimization*. arXiv: 1710.09412 [cs.LG].
- Zhang, Huakui et al. (2022). ‘Multimodal Topic Modeling by Exploring Characteristics of Short Text Social Media’. In: *IEEE Transactions on Multimedia*, pp. 1–1. DOI: 10.1109/TMM.2022.3147064.

Zheng, Yin, Yu-Jin Zhang and Hugo Larochelle (2014). ‘Topic Modeling of Multimodal Data: An Autoregressive Approach’. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1370–1377. DOI: 10.1109/CVPR.2014.178.

Zosa, Elaine and Lidia Pivovarova (Oct. 2022). ‘Multilingual and Multimodal Topic Modelling with Pretrained Embeddings’. In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 4037–4048. URL: <https://aclanthology.org/2022.coling-1.355>.



 **NTNU**

Norwegian University of
Science and Technology