Trade Offs Between Large Language Models and Traditional Statistical Algorithms for Topic Modeling

Marco Schweiss

Contents

List of Figures

List of Tables

List of Abbreviations

API - Application Programming Interface

BERT - Bidirectional Encoder Representations from Transformers

BoW - Bag of Words

GPT - Generative Pre-trained Transformer

HGTM - Hashtag Graph-based Topic Model

HDP - Hierarchical Dirichlet Process

IGTM - Image-Regulated Graph Topic Model

LDA - Latent Dirichlet Allocation

LLM - Large Language Model

LSA - Latent Semantic Analysis

NLP - Natural Language Processing

NMF - Nonnegative Matrix Factorization

NPMI - Normalized Pointwise Mutual Information

pLSA - Probabilistic Latent Semantic Analysis

PMI - Pointwise Mutual Information

SeqLDA - Sequential Latent Dirichlet Allocation

sLDA - Supervised Latent Dirichlet Allocation

TF - Term Frequency

TF-IDF - Term Frequency Inverse Document Frequency

1 Introduction

Topic modeling is an influential statistical methodology employed to identify dominant themes and reveal the inherent semantic framework within extensive text collections. Its applications span a wide range of fields, including information retrieval, sentiment analysis, trend discovery, the biomedical domain, software engineering, and the analysis of social networks. By organizing and making sense of extensive and unstructured data, topic modeling enables researchers to identify patterns that might otherwise remain hidden. The methodology acts as a conduit for illustrating the vast quantities of data produced by advancements in computer and web technologies, which is done by transforming important latent variables and prominent features inside text corpora into a more accessible low-dimensional representation. (Churchill & Singh, 2022)

1.1 Foundational Methods and Limitations

Classical statistical topic modeling methods, such as latent semantic analysis (LSA), probabilistic latent semantic analysis (PLSA), and latent Dirichlet allocation (LDA), are considered to be the foundation of topic modeling and have been the focus of research for around three decades. These methods have undergone extensive research, evaluation, and refinement, making them effective tools for identifying prevalent themes in text corpora. , (Alghamdi & Alfalqi, 2015), (Abdelrazek et al., 2023)

Despite the proven effectiveness of classical methods, there is an ongoing drive within the research community to develop newer and more sophisticated topic modeling techniques. This stems from the limitations of classical models regarding their ability to handle the characteristics of more recent and novel text data types. The increasing complexities of these different types of data, including short texts with few words (Qiang et al., 2022), (Likhitha et al., 2019), multi-modal information (Churchill & Singh, 2022), and the need for incorporating external knowledge sources (Yang et al., 2020), (M. Xu et al., 2017), (Xie et al., 2015) have spurred the exploration of advanced approaches.

1.2 Branching Out

One prominent direction in this quest for improvement is the rise of deep learning-based topic models. By leveraging neural networks, these models aim to learn more complex and flexible representations of text data. (H. Zhao et al., 2021) Additionally, researchers have been exploring specialized topic models tailored to handle dynamic and temporal aspects of data, allowing for insights into evolving trends and themes over time. (Hong et al., 2011), (Delvin & Hady, 2022)

Furthermore, the incorporation of external knowledge sources into topic modeling methods has shown promise for enriching the semantic understanding of the underlying topics. This integration empowers topic models to make more informed and contextually relevant inferences, thus enhancing their utility in real-world applications. (X. Zhao et al., 2021), X. Xu et al. (2022)

Moreover, the emergence of various topic models, such as Author-Topic Models (Rosen-Zvi et al.,

2012), Structured Topic Models (Hanna M., 2008), and Neural Topic Models (H. Zhao et al., 2021), indicates a diversification of approaches to address specific challenges and limitations faced by classical methods. These newer models strive to handle domain-specific complexities, adapt to varying data types, and improve the interpretability and robustness of topic modeling results.

While classical statistical topic models remain commonly used methods for topic modeling, the ongoing exploration and adoption of newer approaches signify a collective effort to overcome limitations and better align with the demands of modern data sources and applications.(Abdelrazek et al., 2023) As scientific advances in unsupervised machine learning continue to drive progress in the field, topic models are poised to play a crucial role in summarizing and understanding the vast and ever-expanding digitized archives of information.

1.3 Paradigm Shift

As we delve into the heart of this transition, a crucial question emerges: How do these newly emerging artificial intelligence (AI)-based methods, particularly the most recent LLMs like GPT-3 and his successors, compare with the well-established traditional statistical algorithms in the realm of topic modeling?

These models, characterized by their impressive pre-training and multitasking capabilities, resulted in fundamentally new approaches to natural language processing (NLP). Their innate ability to recognize a vast number of textual patterns and infer semantics allows them to operate effectively as unsupervised problem solvers. This means that, unlike traditional algorithms that often require laborious fine-tuning and domain-specific adaptations, LLMs have the innate capability to adapt to diverse use cases with minimal or no additional training. (Radford et al., 2018)

Given the general push towards AI methodologies in the topic modeling space, the objective of this thesis is to conduct a systematic and in-depth analysis of the trade-offs between state-of-the-art LLMs and traditional statistical algorithms in the realm of topic modeling. Specifically, the thesis aims to address the following questions:

1. What are the distinctive characteristics of LLMs and traditional statistical algorithms when employed for topic modeling?
2. How do these approaches perform in terms of computational intensity and topic quality, measured by quantitative and qualitative means?
3. What are the implications of choosing one approach over the other based on the specific goals, dataset characteristics, and available resources?

By examining these questions, the thesis will support researchers and practitioners in making informed decisions when selecting the most suitable approach for their specific requirements. The

research not only contributes to the academic discourse on topic modeling but also serves as a practical guide for navigating the intricate landscape of LLMs and traditional statistical algorithms in the context of topic modeling.

1.4 Thesis Structure and Methodology

The beginning of the thesis consists of a comprehensive literature review, which serves as the foundational element. This section offers an overview of prior research endeavors, encompassing both traditional statistical algorithms and the utilization of LLMs for topic modeling. The literature review entails the identification of key studies, findings, methodologies, and models that have contributed to this domain. By analyzing the existing literature, this section aims to identify gaps and limitations that will be the focal point of this thesis.

The review will start with an in-depth exploration of the evolution of topic modeling algorithms from traditional statistical algorithms to neural models. By examining the advancements made in the field, we can gain a deeper understanding of the capabilities and limitations of these algorithms, ultimately gaining a better insight into their potential applications. Following, there is a detailed analysis of topic model evaluation metrics with a discussion of their advantages and drawbacks. In the realm of topic modeling, evaluating the performance of models is crucial for ensuring accurate and meaningful results. By understanding these metrics, researchers and practitioners can make informed decisions regarding the effectiveness or limitations of their topic modeling results.

The following section contains a comparison of algebraic, probabilistic, and LLM-based topic models. The basis for the analysis is a review of multiple studies dedicated to the respective topic models. This analysis is framed around multiple factors, including computational costs, topic quality, preferred document types, and robustness. Additionally, we will look at the granular details of the methodologies, parameters, and datasets that were employed in the reviewed papers. This serves as a basis to detect commonalities, trends, and potential approaches for the experiment that is done later on in the thesis. Lastly, the findings are summarized, and different scenarios are considered and discussed, determining which topic model type is better suited for a specific situation.

Subsequently, the thesis transitions into a thorough examination of LLMs, starting with OpenAI's GPT-3 model. Within this section, the workings of the newest generation of LLMs are discussed, and the relevance of these models in the realm of topic modeling is outlined. Given the relatively recent emergence of this type of LLM, we will use our discoveries regarding the field of topic modeling as a whole and our understanding of the usage of the LLM to derive a suitable methodology for topic modeling that extends the current approaches.

The thesis then progresses into an experimental section, where traditional statistical algorithms are compared to the GPT-3.5 model underlying ChatGPT regarding their performance on different doc-

ument types. The results are then subjected to quantitative measures for comparison. Challenges encountered during the experiment are addressed, and the strategies employed to mitigate them are expanded upon. Continuing the analysis, the subsequent chapter focuses on a qualitative assessment with the GPT-4 model, which is designed to evaluate the quality of topic modeling results derived from the aforementioned experiment. The outcomes of the qualitative assessment are analyzed to determine the preferred method for each of the different document types and to evaluate if automated topic model metrics and qualitative assessment coincide.

Finally, the thesis concludes with a summarized overview of the entire thesis, including the research conducted and the outcomes obtained. The implications of the findings are discussed, and potential improvements and developments within the field are explored.

2 Literature Review

To establish a solid theoretical foundation and further evaluate the strengths and weaknesses of the individual topic modeling approaches, this literature review first covers the evolution of the techniques used in the field. Thereby, we will examine selected papers on key techniques regarding their objectives, limitations and findings on the performance of the techniques on different document types. We will move forward in a problem-oriented fashion and derive an understanding of why the field of topic modeling has developed into its current form. Regarding the selection criteria for the literature, there was an emphasis on including important papers that defined the field as a whole and also studies that contributed novel but more niche topic models that display the diversity of the field.

2.1 From Algebraic to Probabilistic Models

We start the review with LSA, which can be considered the precursor to all other topic models, although the term topic was not explicitly used during the time of creating the model. LSA was developed by Deerwester and colleagues with the main goal of providing an innovative approach to automatic indexing and retrieval that overcomes the limitations imposed by the existing techniques at the time, which were dependent on term-matching. The authors' objective was to address the challenge faced by users in retrieving documents based on conceptual content, as individual words do not offer reliable indications of a document's conceptual topic or meaning.

Although the paper reported that the proposed approach of LSA was superior to simple term matching, the authors themselves considered the result only "modestly encouraging" as the study faced several drawbacks and limitations. This includes faults in the methodology and issues with the used datasets. (Deerwester et al., 1990) Nonetheless, this paper marks the beginning of what would become the field of topic modeling.

Building upon the LSA model, we will next look at pLSA which was first employed by Hoffman. While LSA is based on matrix factorization, pLSA works with a statistical generative model, which marks the

transition from non-probabilistic to probabilistic models. The pLSA model was developed to address some of the limitations of LSA, like the unsatisfactory statistical foundation.

The generative model underlying pLSA allows for dealing with polysemous words, domain-specific synonymy, as well as distinguishing between different types of word usage. Additionally, due to the aforementioned theoretical advantages of pLSA compared to LSA, the study found that pLSA outperformed both LSA and the older term matching approach on all types of document collections that were tested. (Hofmann, 1999)

2.2 The First Topic Model and Extensions

The next model we will examine is called latent Dirichlet allocation (LDA), with Dirichlet referring to the probability distribution that is used in the LDA model. It was developed by Blei, Ng, and Jordan and is certainly the most influential and the first ”real” topic model. The paper in which the model is showcased is also the first that explicitly uses the term topic.

pLSA suffers from several problems, including overfitting due to the linear growth of parameters with corpus size, difficulty in assigning probabilities to new documents outside the training set, and limited flexibility due to several constraints. Therefore, LDA was proposed as an improvement over pLSA. The authors report that LDA is a powerful and flexible model that can effectively capture the latent structure of large collections of documents and provide accurate predictions for various tasks, outperforming the older pLSA approach.(Blei et al., 2003) Nonetheless, the LDA model also has some limitations, which consequently lead to extensions of the original model.

2.2.1 Parameter Selection and Topic Correlation

An important aspect of the LDA model is that the number of topics that are expected in the documents must be defined manually.(Teh et al., 2004) Finding the best value for the parameter happens mostly by testing different values and evaluating the resulting topics. (Churchill & Singh, 2022)

A model that alleviates this issue is called the Hierarchical Dirichlet Process (HDP). The main advantage of HDP over LDA is that HDP does not require the specification of the number of topics beforehand, unlike traditional LDA models. Instead, the number of topics is inferred from the data, which can be useful when the number of topics is unknown or may change over time. The authors of the paper that proposed the HDP report that the model performs on par with the optimal LDA model for the documents used in the study, with the benefit of not having to define the number of topics beforehand. (Teh et al., 2004)

Another limitation of LDA is that the model assumes that the topics in a document collection are independent of each other, which is not always the case in real-world scenarios. The Correlated Topic Model (CTM) is designed to address this limitation in the LDA model by using a more flexible distribution, namely the logistic normal distribution. This change allows for correlation between the

topic proportions and provides a more realistic model of the latent topic structure.

According to the authors of the model, the CTM was found to give a better fit than the LDA on the documents employed in the study. Furthermore, the CTM was found to provide a natural way of visualizing and exploring unstructured collections of textual data. (Blei & Lafferty, 2005)

2.2.2 Bag-of-Words Assumption and Target variables

LDA is based on the bag-of-words assumption, which means that the order of words in a document can be neglected. The authors of the LDA already noted in the original paper that the bag-of-words assumption allows words that should be generated by the same topic to be allocated to several different topics. (Blei et al., 2003) The problem with ignoring the order of words in a document is that it can result in less meaningful inferred topics.

To combat this issue, Wallach proposed a hierarchical Bayesian model that integrates bi-grams into the topic modeling process. The authors report that the bi-gram topic model outperforms LDA on two data sets in terms of predictive accuracy. Additionally, the inferred topics are less dominated by function words than are topics discovered using LDA, potentially making them more meaningful. Function words are words that serve a grammatical or structural role in a sentence rather than conveying content or meaning, for example, ”the”, ”a”, ”in”, ”on”, etc. (Hanna M. Wallach, 2006)

LDA is an unsupervised topic model algorithm, meaning that only the words in the documents are modeled. In general, the original LDA is not really suitable for prediction tasks because it is not designed to infer topics that are predictive of a response variable. For example, to predict a movie rating based on the topics in a review. This stems from the dimensionality-reducing component of LDA, which results in topics that correspond to dominant structures in the corpus instead of being useful for predictions of certain target variables.

To solve this problem, Blei and McAullife developed a supervised variant of the LDA model called sLDA. The initial sLDA model is trained with a labeled dataset using a maximum likelihood approach. After training, the sLDA model can then be used to predict the response variable for new, unlabeled documents based on their inferred topic distributions.

Testing the sLDA model on two prediction problems revealed a better performance of the sLDA compared to using standard LDA topics with regression afterward. (Mcauliffe & Blei, 2007) However, the fact that sLDA is a supervised model means that it requires labeled data for training, which can be a limitation in some cases where labeled data is not available or is difficult to obtain.

2.2.3 Topic Evolution

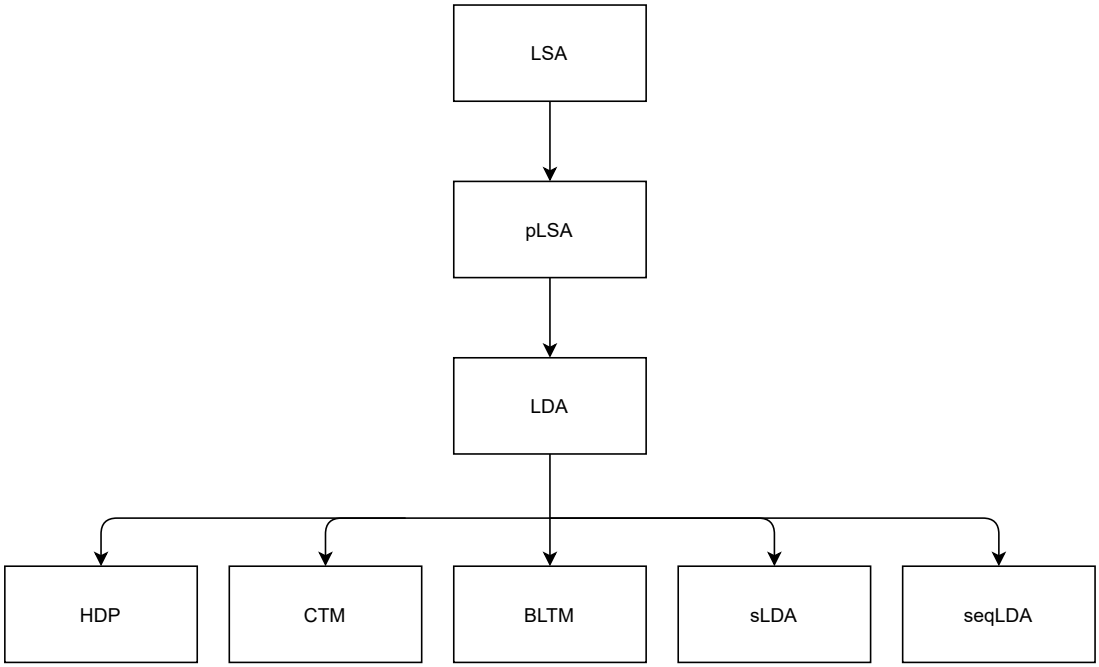
Another limitation of LDA that results from the bag-of-words assumption is that the model ignores the sequential structure of each document. The sequential LDA (SeqLDA) model aims to address this issue by accounting for the position of each segment within a document to explore how topics evolve

within a document.

The SeqLDA model accounts for the evolution of topics in a document by explicitly modeling the underlying document structure, specifically the individual segments. Thereby, the topic distribution of each segment depends on that of its preceding segment, and the progressive topical dependency is captured.

According to the authors, the SeqLDA outperforms LDA and generates a superior sequential structure of the topics in their experiment on a collection of books. (Du et al., 2012) To summarize the development of the different statistical models that have been covered until now, the following figure illustrates the order of development:

Figure 1: Development of traditional topic models



Source: Own results

2.3 Novel Approaches

After looking at the classical statistical models and more modern derivations of them, specifically variants of the prominent LDA model, we will now look into novel approaches to topic modeling.

2.3.1 NMF

Non-negative Matrix Factorization (NMF) is, like LSA, a non-probabilistic method of topic modeling. Thereby NMF is an approach that approximates a non-negative matrix by computing the product of two low-rank non-negative matrices. Its capability to generate results that are meaningful regarding their semantics and that can be easily interpreted in clustering scenarios has led to the widespread adoption of NMF as both a clustering method for document data and a technique for topic modeling. (Kuang et al., 2015)

Compared to other methods, NMF may be a superior choice when noisy datasets are used for topic modeling. This is because the utilization of pure dimensionality reduction approaches like NMF, which is based on matrix factorization, allows eliminating noise and extracting features from sparsely occupied high-dimensional spaces. (Churchill & Singh, 2022) Nonetheless, like many other topic modeling algorithms, several variants of NMF have been proposed over the years to account for its shortcomings. For example, semi-supervised NMF models that differ from standard NMF in that they incorporate additional information in the factorization process. (Haddock et al., 2020)

NMF has gained significant popularity in the field of topic modeling for analyzing extensive documents. However, the topics generated by NMF often tend to be overly general and redundant, lacking in minor yet potentially valuable information for users. To address this issue, Suh and colleagues presented an ensemble model of NMF that aims to uncover localized high-quality topics. The approach involves utilizing an ensemble model to iteratively conduct NMF using a residual matrix derived from previous stages, resulting in a series of topic sets. (S. Suh et al., 2016)

2.3.2 Graph-based models

Another, more recent approach to topic modeling is graph-based methodologies. In a paper by Wang and colleagues, the authors introduced a topic model called the Hashtag Graph-based Topic Model (HGTM) for handling semi-structured tweets. By leveraging the relationships between hashtags, the HGTM establishes semantic associations between words. During the time of publication, the attention towards mining topics on Twitter has been steadily increasing. Nevertheless, the concise and informal nature of tweets results in a sparse vector representation that encompasses a vast vocabulary, which leads to conventional topic models such as LDA frequently falling short of generating quality topics. The HGTM has been demonstrated to be highly effective in uncovering a greater number of distinct and cohesive topics. Additionally, the model has exhibited robust capability in managing sparseness and noise within tweets. (Y. Wang et al., 2014)

In an effort to improve news data analysis, Zhang proposed constructing a keyword-keyword network using a graph structure to improve upon current technologies that primarily rely on LDA-like models. The author argues that these models fail to explore the intricate semantic connections among a set of topic-related keywords. (Lidan, 2022) Another limitation of traditional topic models in discovering latent topics from cross-media data becomes apparent when text is combined with additional information such as geo-information, user-annotated tags, pictures, and videos. To address this issue, a graph-based model called the Image-Regulated Graph Topic Model (IGTM) was introduced. By integrating relational information among images into the modeling process, IGTM successfully uncovers higher-quality underlying topics. (Z. Wang et al., 2015)

2.4 Towards Transformer Models

To discuss another class of novel approaches, we will now look into transformer-based topic modeling, which started with the advent of Word2Vec. The introduction of Word2Vec brought about a notable transformation in the realm of topic modeling and NLP. Word2Vec revolutionized the creation of word embeddings by introducing a novel approach that was characterized by its efficiency and precision. Moreover, it demonstrated the potential of utilizing these word vectors to identify words with similar semantic meanings.

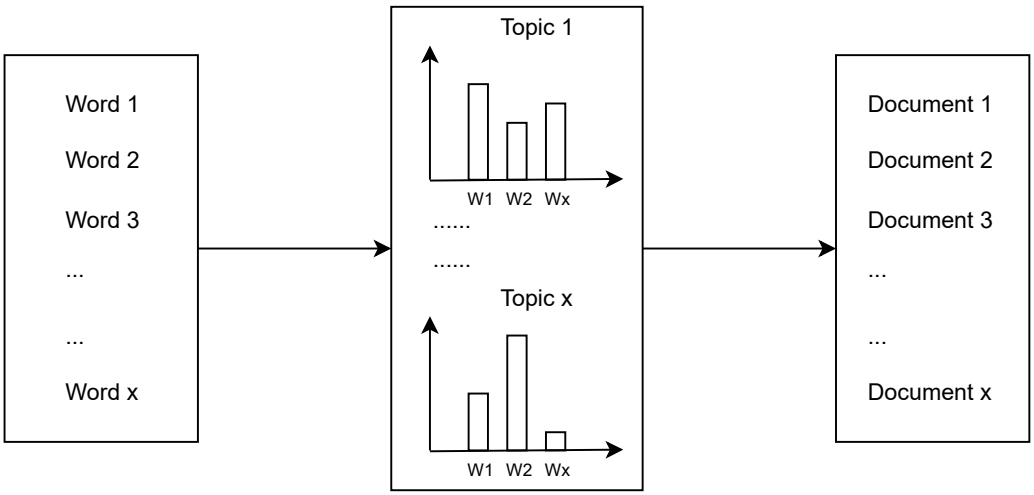
Clustering these embeddings basically provides the same insights into the semantic structure as traditional topic models like LDA. (Thompson & Mimno, 2020) This simplicity and effectiveness of word embeddings made them the most popular form of modern NLP model that has been incorporated into topic models. Word embeddings have become a fundamental component of many modern NLP models, including BERT, which is a LLM that is based on word embeddings. (Churchill & Singh, 2022)

The novel approach of clustering word embeddings for topic modeling was proposed by Thompson and Mimno who used LLMs like BERT, GPT-2, and RoBERTa. The models were employed by the authors to generate word embeddings, which were subsequently subjected to clustering using a k-means algorithm. It was observed that the resultant word clusters exhibit comparable characteristics to those produced by an LDA model. In addition, a comparison was made between the performance of these cluster models and LDA topic models. The results revealed that the cluster models can achieve similar or even better performance than their LDA counterparts. (Thompson & Mimno, 2020)

A similar approach was used by Grootendorst who proposed the BERTopic model. The topic modeling starts by representing each document as an embedding, which is done with the SBERT model, a variation of the original BERT. Following that, there is a dimensionality reduction step to optimize the clustering process. After clustering, the topics are extracted using a variation of TF-IDF that is designed to model the importance of words in clusters of documents rather than individual documents. The paper reports that BERTopic generally performs well across all used datasets. (Grootendorst, 2022)

To conclude the review of topic modeling development, the figures below illustrate the different general procedures for either LDA-based models or models using word embeddings. Traditional probabilistic models assume that documents are comprised of latent topics, which in turn consist of a distribution of certain words. (Blei, 2012) This is illustrated in figure 2, which depicts the thought process behind LDA. The left column represents the vocabulary of the entire corpus, which is used to model our topics. The middle column contains the topics, which consist of the vocabulary and the corresponding probabilities of individual words belonging to each topic. This is visualized using bar graphs. Finally, the right column contains the corpus, which is a collection of documents where each document is assumed to consist of various topics.

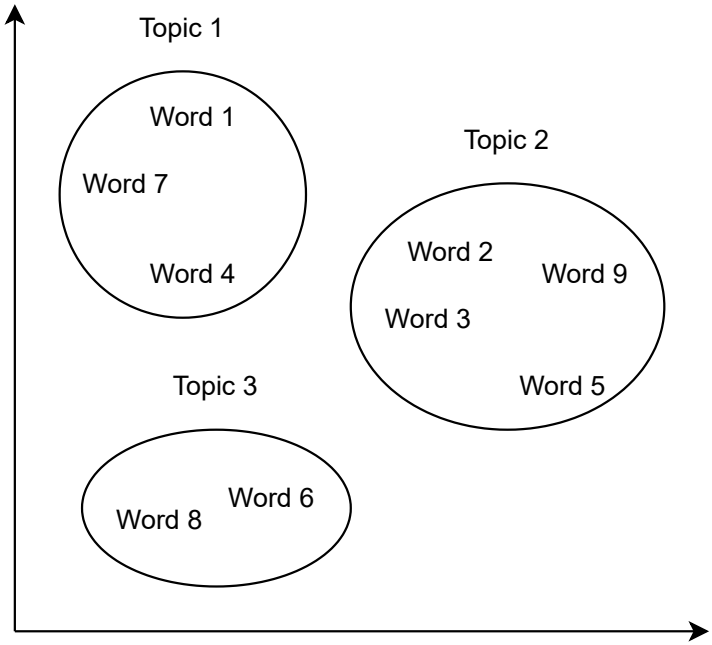
Figure 2: Topic modeling process for the LDA model



Source: Adapted from: Blei (2012), p. 78.

On the other hand, approaches based on word embeddings use contextualized word representations to capture the semantic relationships between words in a corpus and group them into clusters based on their similarity, which happens to produce similar outputs as the LDA model. (Thompson & Mimno, 2020)

Figure 3: Topic modeling process for word embedding models



Source: Adapted from: Li et al. (2019), p. 691.

2.5 Performance Metrics

After examining the general direction that the field of topic modeling has taken since its inception, we will now look into the different performance metrics for evaluating topic models. Specifically, we will examine what evaluation metrics exist and their respective strengths and weaknesses. By looking at

the research done, we can uncover a variety of different evaluation metrics, for example, perplexity (Blei et al., 2003), coherence (David et al., 2010), stability (de Waal & Barnard, 2008), coverage, and, as a qualitative method, human evaluation. Topics generated by algorithms are considered beneficial if they can be comprehended by humans. Furthermore, human assessments hold significance in topic modeling as they enable the development and validation of automated evaluation metrics (Matthews, 2019). Despite the progress made in topic modeling through automated methods, human assessment continues to be crucial in evaluating and refining these techniques. This is because it serves as a benchmark to ensure that the generated topics are in line with human understanding and expectations.

What all methods have in common is that there is no one perfect evaluation metric that fits every type of use case. The metrics employed thus far present a varied depiction, rendering the verification of the topic modeling results challenging. In sum, both perfectly selecting a suitable algorithm and assessing the outcomes continue to be unresolved matters. Ideally, a set of different metrics is employed in addition to qualitative human evaluation to get a comprehensive understanding of a model's performance. (Rüdiger et al., 2022), (Churchill & Singh, 2022).

2.5.1 Perplexity

The capability of the model to generate the documents in the corpus based on the learned topics is measured by perplexity. Perplexity evaluates the model's predictive power, thus indicating how well the model explains the data. If the information gained from learning the outcome of a random variable is minimal, it implies that the model is perplexed. (Abdelrazek et al., 2023) The usage of perplexity as a metric for topic models has a long history, for example, the authors of the original pLSA (Hofmann, 1999) or the LDA used it for evaluation in their study, where they argued that a lower perplexity score indicates better generalization performance of the model. (Blei et al., 2003)

In order to examine the connection between perplexity and information retrieval performance, Azzopardi and colleagues conducted an empirical study using the pLSA model. The findings demonstrated a predictable relationship between topic model perplexity and precision-recall performance, which was observed across multiple corpora. (Azzopardi et al., 2003) On the other hand, another paper by Blei notes that the held-out accuracy that perplexity represents may not necessarily correspond to good topic interpretation, which is an important goal of topic modeling. Therefore, there is a need to develop evaluation methods that match how the algorithms are ultimately used. (Blei, 2012)

The critique of perplexity as a topic modeling evaluation metric is thereby common in the scientific literature. According to a study conducted on the evaluation of topic models, it was found that perplexity fails to measure the coherence of topics. The study suggests that topic model evaluation should instead prioritize task-specific performance. Interestingly, topic models that achieve higher scores in held-out likelihood may actually generate less semantically meaningful topics. While perplexity can

be useful in assessing the predictive performance of topic models, it does not address the exploratory objectives of topic modeling. (Chang et al., 2009)

Another paper that compared perplexity to another evaluation metric called topic stability reports that perplexity is plagued by several deficiencies. One such issue is that perplexity is contingent upon the vocabulary size being modeled, rendering it unsuitable for comparing models that employ distinct input feature sets or operate in different languages. (de Waal & Barnard, 2008)

2.5.2 Topic Coherence

Given that the perplexity metric favors different models than human judgment, there was a significant shift towards other metrics, one of them being topic coherence. (Hoyle et al., 2021) Topic coherence refers to the degree to which the words or concepts in a given topic or set of topics are logically connected and support each other. In other words, a coherent topic is one where the words or concepts are related and make sense together, while an incoherent topic is one where the words or concepts are unrelated or do not fit together logically. Thereby, topic coherence is not one single metric but a whole group of metrics, for example, pointwise mutual information (PMI), normalized pointwise mutual information (NPMI), or cosine similarity. (Röder et al., 2015)

Automated topic coherence measures were introduced by Newman and his team. They utilized resources such as WordNet, Wikipedia, the Google search engine, and previous research on lexical similarity and relatedness. By comparing human evaluations of learned topics from two different datasets, the study provided evidence that a straightforward co-occurrence measure that relied on PMI yielded results for the task that were highly similar to the level of agreement among human evaluators. (David et al., 2010) In another paper about topic coherence measures done with LDA and LSA as topic models, the researchers found that the automatic topic coherence evaluation aligns with human evaluations.

In another paper on topic quality metrics, the author reports that automated evaluation of topic quality remains an important unsolved problem in topic modeling that represents a major obstacle to the development of new topic models and that human judgment can be considered the gold standard in this field. (Nikolenko, 2016) However, automatically assessing the cohesiveness of the identified topics poses a challenge as an unsupervised task and does not ensure the interpretability of the topic model. To gain further insights into the performance of various coherence metrics in topic modeling, Campagnolo and his team conducted an analysis to determine their sensitivity. Sensitivity was measured by examining how these metrics behaved when applied to both well-formed and noisy topics, where noisy topics are those containing irrelevant words. To validate the quality of the topics, a qualitative survey was conducted with more than 60 participants, providing a benchmark for comparison.

The analysis conducted by the researchers reveals that specific metrics exhibit a higher susceptibility

to noise, thereby validating their suitability in situations where users aim to emphasize topics containing unrelated terms. Conversely, alternative metrics display greater resistance to corrupt data and are less affected by noisy information. These metrics can be employed when users seek to identify more authentic topics among those uncovered. (Campagnolo et al., 2022)

On the other hand, Hoyle and colleagues doubt the usefulness of automatic topic coherence metrics in general. Their study, which aimed to investigate the validity of automated coherence measures for evaluating topic models, reports that these metrics have limitations and may not be fully reliable in evaluating topic models. The authors argue that coherence measures designed for older models may be incompatible with newer models, and automated evaluations declare a winning model when corresponding human evaluations do not. (Hoyle et al., 2021)

2.5.3 Topic Coverage and Diversity

The evaluation of topic coverage relies on a collection of reference topics and coverage measures that assess the extent to which the model topics align with the reference topics. The reference topics signify the subjects of interest that topic models should uncover. Once a compilation of reference topics is established, the coverage of these topics by a specific topic model instance refers to the percentage of reference topics that are matched by the model's topics. A reference topic is deemed covered if one or more corresponding topics generated by the model exist. (Korenčić et al., 2021) In one of the initial publications that introduced this metric, Chuang et al. conducted a comprehensive analysis by comparing 10,000 variations of topic models with 200 domain concepts provided by experts. Their findings showcased the metric's ability to guide decisions on the choice of topic model and the model's parameters. (Chuang et al., 2013)

In an attempt to enhance the automation of topic coverage measurement, Korencic and his team introduced an unsupervised method. Their approach utilizes topic distance as a criterion for matching topics and incorporates a range of coverage scores calculated for various distances. The paper includes the design and evaluation of coverage metrics, as well as coverage experiments conducted on two datasets. The authors demonstrate that this measure exhibits a strong rank correlation with a supervised measure. Furthermore, the unsupervised measure can be easily applied to new datasets and utilized for model selection and evaluation through ranking a set of topic models. (Korenčić et al., 2021)

A metric that is related to coverage although different is topic diversity. Dieng and his team define diversity as the proportion of distinct words among the top 25 words across all topics. An ideal topic model should produce diverse topics and achieve a high score on this measure. Conversely, a low score suggests the presence of repetitive topics, indicating that the model was unable to effectively separate the themes within the corpus. (Dieng et al., 2020) Good topic diversity increases the likelihood of encompassing all the themes present in the corpus, which holds significance in downstream

applications like text summarization and classification. Topic diversity as a metric has a possible application in parameter selection because it may indicate the optimal number of topics to be modeled. If many topics are selected, there is a risk of having similar topics with overlapping keywords. Conversely, if a few topics are chosen, the resulting topics may be broad and difficult to interpret. (Abdelrazek et al., 2023)

2.5.4 Topic Stability

The last metric that we will look at is topic stability. As for topic coherence, various metrics can be used to represent the stability of a topic model. However, the fundamental approach involves assessing the similarity between topics in multiple iterations of topic inferences. Greater model stability is achieved when similar topics are consistently generated across different iterations. (Abdelrazek et al., 2023) Stability across topics was first mentioned by Steyvers who argued that in topic modeling, there are cases where it is beneficial to concentrate on a single specific topic or theme in order to better understand each individual topic. In such situations, it is important to determine which topics consistently appear across different inferences and which topics are unique and specific to a particular inference. (Steyvers & Griffiths, 2007)

In a paper that investigated the improved suitability of topic stability as an evaluation metric compared to perplexity, the authors argue that one of the key attributes of a useful topic model is that it should model corpus contents in a stable fashion. That is, useful topics are those that persist despite changes in input representation or model parameters. (de Waal & Barnard, 2008) Kherwa and Bansal observed that the majority of authors in topic model papers tend to overlook this matter, opting instead for a single random initialization and asserting the outcome of the topic modeling experiment as conclusive. (Kherwa & Bansal, 2019)

Topic stability can thereby be used to control for a variety of different factors in topic modeling. For example, in their investigation into the robustness of topics in the face of noisy datasets, the authors assert that the inclusion of erroneous or noisy texts in corpora has the potential to undermine topic stability. Thus, understanding how well a topic modeling algorithm performs when confronted with noisy data becomes imperative. The study reveals that different types of textual noise can exert varying impacts on the stability of topic models. (Su et al., 2015) Greene and colleagues discuss in a separate study on topic stability that, despite the various topic modeling algorithms proposed, a common obstacle to effectively utilizing these techniques is the choice of an optimal number of topics for a given corpus. They argue that a model with the right number of topics will be more resistant to changes in the data and consequently demonstrate improved topic stability. (Greene et al., 2014)

2.6 Comparative Performance Analysis

After examining the metrics used in the field of topic modeling, we will now look at the characteristics of how topic models behave in terms of the following points:

- **Scalability:** What are the computational costs of a model, and what size of dataset is optimal?
- **Topic Quality:** How do the models compare to each other regarding the quality of the resulting topics measured by various metrics?
- **Document Characteristics:** What type of documents does a model perform best on, and are there documents that do not work well?
- **Robustness:** What influence do preprocessing steps and parameter selection have on performance?

The focus here will be on algebraic, probabilistic, and transformer-based models representing each of the major categories of topic modeling approaches. Additionally, we will look at the methodologies, metrics, datasets, and data processing steps of each respective study.

2.6.1 LSA and LDA

In the initial paper, the authors of LSA report that the method is scalable and can be applied to large collections of texts, given the economical representation of the documents. (Deerwester et al., 1990) Another early study done on LSA mentions that in order to conduct the matrix decomposition of the corpus, a substantial volume of text is necessary. The inclusion of more text improves the quality of the model's outputs by offering numerous contexts where words can co-occur with one another. Thereby, a minimum of 200 contexts, such as sentences or paragraphs, are typically required. (Foltz, 1996)

Moving away from older papers on LSA that did not really focus on topic modeling, we will now look at a more recent study conducted by Zengul et al. The objective of the study was to compare three different topic modeling methods, namely LSA, LDA, and Top2Vec, using a COVID-19 textual data corpus. The study aimed to provide guidance for researchers interested in using topic modeling methodologies and to help them determine which methodology to use. Interestingly, the paper does not try to declare a superior topic model but compares the output of the three models in terms of similarity. The outcome of the experiment reveals that LDA topics have a high correlation with Top2Vec topics, followed by LDA and LSA, and lastly LSA to Top2Vec.

Regarding scalability, they found that LSA does not require such high computational resources for the same size of data compared to LDA. The authors argue that although both LSA and LDA yield quality topics if solid data preprocessing is done, LSA should be preferred when computational resources are of concern. The paper reports that topic modeling on a 65.000 abstract dataset is possible with a desktop computer when LSA is used but requires significant computing resources when LDA is employed. The same holds true for Top2Vec which is a word embedding-based methodology that also requires significant computing resources. (Zengul et al., 2023)

Regarding scalability, they found that LSA does not require such high computational resources for the same size of data as LDA. The authors argue that although both LSA and LDA yield quality topics if solid data preprocessing is done, LSA should be preferred when computational resources are of concern. The paper reports that topic modeling on a 65,000 abstract dataset is possible with a desktop computer when LSA is used but requires significant computing resources when LDA is employed. The same holds true for Top2Vec, which is a word embedding-based methodology that also requires significant computing resources. (Zengul et al., 2023)

In a paper that compared the effectiveness of LDA and LSA in the context of a content-based movie recommendation system, we can derive another interesting insight into the behavior of these models regarding computational costs. It is mentioned that LDA had a lower computational cost than LSA, but it is important to note that 500 topics were modeled with LSA, whereas LDA was used to model only 50 topics. According to the authors, the different topic numbers were chosen because LSA performs better for a greater number of topics. (Bergamaschi & Po, 2015) Another similar study that compares LSA to LDA on an e-book dataset reports that in terms of topic coherence, LDA outperforms LSA. The highest coherence value was achieved by the LDA model when 20 topics were used, whereas the LSA model performed best with 10 topics. This is in direct contrast with what was mentioned in the previous study. Nonetheless, what both papers have in common is that they consider the pre-processing of the text data to be vital for good topic quality (Mohammed & Al-augby, 2020)

Another paper that compared LSA to LDA on a scientific abstract dataset further delivered evidence that LDA performs better than LSA regarding topic coherence. Although, oddly enough, because a standard implementation of LSA and LDA was used in this study, LDA outperformed LSA regarding runtime for every number of topics to be modeled. What coincides with other papers is the fact that different preprocessing steps of the text corpus have an influence on the resulting topics. In this case, there was an altered coherence score. (Bellaour et al., 2021)

Much like choosing appropriate preprocessing measures, the identification of the best-performing number of topics in LSA poses a significant challenge. This is backed by a paper that introduces LSAView, a system aimed at facilitating interactive exploration of parameter choices for LSA models. The authors argue that determining the appropriate model parameters to employ for various data domains and types of analyses represents one of the most significant challenges when using LSA. (Crossno et al., 2009) In their study, Naili and her team aimed to analyze the impact of these parameters on topic segmentation quality and to determine the most effective ones. They report that their experiments demonstrate that the selection of LSA parameters greatly influences topic segmentation quality. (Naili et al., 2018)

Another study that analyzed the performance of LSA on a descriptive answer corpus reports that LSA does not perform well on noisy input data. Furthermore, it was noted that various factors, such as

corpus preprocessing, the generation of the term-document matrix with and without a term weighting function, and the selection of dimensionality, significantly influence the performance of LSA. (Kaur & Kumar, 2019) Evangelopoulos and his colleagues expand widely on this and also provide evidence in support of these findings, suggesting that although LSA has wide-ranging applications, researchers must exercise thoughtful parameter selection and methodological considerations.

The literature extensively discusses the empirical examination of selecting a suitable number of latent semantic dimensions; however, no conclusive findings have been established. Another unresolved concern is term selection, which is necessary to ensure computational efficiency and prevent over-fitting in the semantic space. One commonly used method is to eliminate terms that have a low frequency across the entire set of documents, known as frequency filtering. The authors emphasize the importance of the vocabulary of terms employed in LSA in influencing the analysis's outcomes.

Also, an important issue is the search for the most effective method of weighting term frequencies. When dealing with article titles or brief text messages, implementing a log-entropy transformation may yield superior outcomes. This is because such texts remain more closely aligned with the outer edges of language structure, where a handful of commonly occurring words can significantly impact meaning. On the other hand, TF-IDF proves more adept at uncovering patterns within the central core of language. It identifies larger clusters of terms that tend to occur together at moderate frequencies. (Evangelopoulos et al., 2012)

The last study dedicated to LSA provides another interesting take on the relationship between LSA and corpus size. In a comparison study of LSA and LDA, Cvitanic and his team argue that LSA cannot accurately predict how people associate words. This happens because of the way LSA represents words, making them seem more similar than they actually are and following rules intrinsic to the method that do not always fit real-world language use. Although these criticisms are important at the word level, they might not matter as much when there is a larger collection of documents to work with. (Cvitanic et al., 2016)

2.6.2 NMF and LDA

Moving on from LSA, we will first look at the behavior of another algebraic topic model, namely NMF. In a comparative analysis of LDA and NMF using multiple short text datasets containing texts with average lengths ranging from 3.46 to 14.34 terms, the authors observed that NMF outperformed LDA in terms of topic generation. Short texts are known for their noise and sparsity, resulting in insufficient information for successful statistical learning with LDA. Conversely, NMF proves to be more effective in dealing with this type of data. (Chen et al., 2019) On the contrary, a comparative analysis of different topic modeling techniques revealed that while the LDA and NMF methods produced higher quality and more coherent topics than other methods on a short text Facebook conversation dataset, the LDA method stood out for its flexibility in providing more meaningful and logically extracted topics,

especially when fewer topics were considered.

Additionally, a measurement of the topic coherence score revealed that reducing the number of topics resulted in a higher coherence score for both LDA and NMF methods. While NMF and LDA exhibit comparable performances, LDA demonstrates greater consistency. However, in terms of runtime comparison between LDA and NMF methods, it was found that LDA was slower. (Albalawi et al., 2020) Another research study conducted on collections of online news articles from different sources and Wikipedia pages discovered that NMF consistently generates more coherent topics compared to LDA, which tends to produce topics that are more general and redundant. The choice of term weighting strategy also significantly influences the results in all cases. The findings from NMF indicate that it might be a better approach for topic modeling in specific corpora, particularly those related to specialized or non-mainstream domains. (O'Callaghan et al., 2015)

On the other hand, a study done on Twitter data reports that NMF is not really suitable for short text data because of its high sparsity. Therefore, the authors propose an extension to the standard NMF topic modeling approach to make it more suitable for short texts. (Athukorala & Mohotti, 2022) Another paper also done on Twitter data found that in terms of perplexity and coherence, LDA outperformed both LSA and NMF in generating topics. (Tijare & Rani, 2020) An additional paper adds further evidence of NMF not being as suitable for short text data. In a study that employs LDA and NMF on text data in the form of short crime reports, the authors report that LDA outperforms NMF in terms of topic coherence. (Pandey & Mohler, 2018)

Regarding the interpretability of topics, another similar study done on tweets mentions that the empirical evidence suggests that both LDA and NMF algorithms are effective in detecting topics, with LDA providing more meaningful interpretations and NMF being the faster option. (Suri & Roy, 2017) Similar to LSA, another paper mentions that NMF is a suitable technique in topic modeling for ingesting large document collections. However, the resulting topics often provide general and redundant information about the documents, lacking potentially meaningful minor details that could be relevant. (Suh et al., 2018)

Another study that adds to the evidence is done by Papadi et al. The research objective of the case study was to compare different topic modeling methods for analyzing conversation transcriptions between customers and agents in a call center, with the aim of improving the efficiency and user satisfaction of customer care services. According to their experiment, LDA generally outperformed NMF in terms of several metrics, including coherence across a variety of topics. (Papadia et al., 2023)

2.6.3 BERTTopic and LDA

We will now move forward and look at the neural models, specifically LLM-based approaches like the already introduced BERTopic. The creator of the original model found that BERTopic is generally

slower than traditional topic modeling techniques such as LDA and NMF. To test the performance, the authors used three datasets, with two of the collections containing a combined number of 18,500 news articles and one consisting of 44,253 tweets, which are short text data. Although it was outperformed regarding runtime by the other approaches, across all datasets, BERTopic consistently achieves high topic coherence and topic diversity scores. Like many others, the author mentions the drawbacks of automated metrics. The assessment of coherence and diversity in a topic can differ between users. Therefore, although these measures can offer some insight into the performance of a model, it is important to acknowledge that they are only indicative. (Grootendorst, 2022)

A comparison between topic models further adds to the evidence that BERTopic outperforms LDA on short-text data in the form of tweets from Twitter. Regarding the quality of the topics, the authors note that BERTopic provided a clear distinction between any identified topics, and the model was able to generate novel insights compared to LDA, which delivered only superficial topics. Despite BERTopics good performance, there were also some drawbacks, like that BERTopic may generate too many topics, which need manual processing afterward, and that there are no objective evaluation metrics, which generally makes it difficult to compare its performance to other topic modeling algorithms. (Egger & Yu, 2022) The generation of too many topics was also mentioned for Top2Vec, which follows a similar methodology than BERTopic (Zengul et al., 2023)

Another study that compared the performance of different BERTopic variants found that the choice of clustering algorithm that is used can have a significant impact on the resulting topics. The study used two datasets, the Course Evaluation Responses (CER) dataset, containing 62,522 short texts, and a subset of the 20 Newsgroups (20NG) dataset, curated to contain only short texts. (de Groot et al., 2022) This implies that the results of BERTopic are susceptible to methodological variables like the choice of clustering algorithm for the embeddings and must be chosen to suit the type of data. This means that, like LDA, which relies on parameter selection, the BERTopic model is also not a one-size-fits-all approach and should be fine-tuned depending on the data that should be modeled.

When confronted with certain data, topic models like LDA and LSA do not perform sufficiently well. Such is the case for the examination of responses to open-ended questions. Traditional topic models are often ineffective in this task as they depend on co-occurrences, which are not commonly found in brief survey replies. According to Xu et al., BERTopic outperformed LDA on a social survey dataset by all criteria employed. The model provided results of quality similar to qualitative analysis in social studies. However, like a lot of other researchers, the authors reported that, as a drawback, BERTopic requires more computational resources than traditional topic models. (X. Xu et al., 2022)

Now that we have discussed all the major topic model types regarding the aforementioned performance criteria, we will look at the details of the reviewed papers. As the last part of the performance review, the following table contains important information about the number of topics that yielded the

best performance, the evaluation criteria, the data and the pre- and post-processing steps that were done for each study.

Table 1: Compilation of topic modeling research

Source	Models	Topic	Evaluation	Data	Pre/Postprocessing
Deerwester, 1990	LSA	100	Precision	Abstracts	Removed words that only occur in one document and 439 common words used by SMART.
Crossno, 2009	LSA	30	Use case specific	News	Not mentioned
Bergamaschi, 2015	LSA, LDA	500, 50	Use case specific	Movieplots	TF-IDF weighting
Greene, 2015	NMF	20	Stability (Agreement score)	News, Wikipedia articles	Removed English stopwords, removed terms occuring in less than 20 documents, TF-IDF term weighting and L2 document length normalization
Naili, 2016	LSA	-	Use case specific	25000 word corpus	Stopword removal, stemming and TF-IDF weighting
Cvitanic, 2016	LSA, LDA	150	Use case specific	Patents	Removed symbols, numbers, mispelled words, stopwords and any words common to 90% or more of the patents
Suri, 2017	NMF, LDA	15	Human Assesment	Tweets, Headlines	Removed URLs and stopwords and used TF-IDF weighting
Pandey, 2018	LSA, NMF, LDA	7	Coherence (UMass)	Short crime reports	Removed English stopwords, common crime related words, words with length less than three and used TF-IDF term weighting
Suh, 2018	NMF, LDA	48	Coherence (PMI), Coverage	News, Emails, Research papers, Tweets	Not mentioned
Kaur, 2019	LSA	3	Use case specific	Paragraphs	Removed words that occur only a single time and TF-IDF weighting
Chen, 2019	LDA, NMF	60-100, 100	Coherence (PMI), Human Assesment	Termgroups, News, QAs, Headlines	Not mentioned
Mohammed, 2020	LSA, LDA	10, 20	Coherence (UMass, Cv)	Books	Removed all numbers, symbols, useless words or letters, punctuation, words with less than 3 characters, English stopwords. Lowercasing and stemming.
Albalawi, 2020	LSA, NMF, LDA	50, 20, 50	Recall, Precision, F-Score	News, Facebook conversations	Removed English stopwords, stemming, lemmatizing, bigram-, trigram generation, and TF-IDF weighting
Tijare, 2020	LSA, NMF, LDA	10 (LDA)	Perplexity, Coherence (Cv, UMass),	Tweets	Removed Twitter handles, punctuations, numbers, special characters, stopwords. Stemming, lemmatizing and TF-IDF weighting

Source	Models	Topic	Evaluation	Data	Pre/Postprocessing
Bellaouar, 2021	LSA, LDA	10, 20	Coherence (UMass, Cv)	Research papers	Removed stopwords, and any word composed of a single character. Lemmatizing, lowercasing and bi-gram extraction
Egger, 2022	NMF, LDA, BERTopic	10, 14, 100	Human Assesment	Tweets	Removed mentions (e.g., @users), hashtags, unknown characters, emojis, stopwords, numbers, and abbreviations. Stemming, lemmatizing, TF-IDF weighting, HDBSCAN clustering
Athukorala, 2022	NMF, LDA	-	F-Score, Coherence (PMI), Human Assessment	Tweets	Removed stopwords, punctuation, emojis, usernames and hash symbols. TF-IDF and IDF term weighting
Grootendorst, 2022	NMF, LDA, BERTopic	10-50	Diversity, Coherence (NPMI)	News, Tweets	Removed punctuation, stopwords, and documents with less than 5 words. Lowercasing and lemmatizing
de Groot, 2022	LDA, BERTopic	5-30	Coherence (NPMI), Diversity (Dieng et al.)	News, Comments	HDBSCAN clustering, k-means clustering
Xu, 2022	LDA, BERTopic	9	Coherence (NPMI), Diversity (Inversed RBO), Human Assessment	Surveys	Not mentioned
Papadia, 2023	NMF, LDA	20	Diversity (Inversed RBO), Similarity (RBO), Coherence (Cv), Classification score	Customer Care Transcripts	Removed stopwords and punctuation. Stemming and lowercasing
Zengul, 2023	LSA, LDA	11	Coherence, Use case specific	Abstracts	Removed Englisch stop words and 1149 unique words. Created bi-grams and five-grams. Lemmatizing, words in the third person are changed to the first person and verbs in the past and future tenses are changed into the present.

Source: Own results

2.6.4 Summary and Evaluation

Now that all the results have been compiled, we will summarize the findings, starting with the algebraic models LSA and NMF. The studies suggest that probably the greatest strength of both topic models is their ability to handle large document collections due to their fast computation. Regarding topic quality, the algebraic models yield mixed results, as they are often outperformed by the other models regarding various metrics. Nonetheless, the limitations become less relevant with larger datasets, where the computational costs are more important than granular topics.

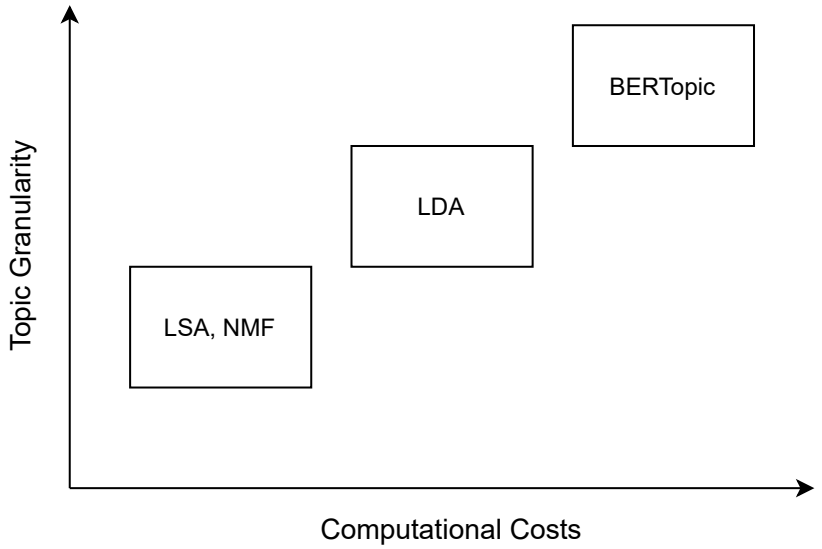
If we compute the average number of topics from the table above that got the best results in the studies, we get around 40 topics for LSA (excluding the 500-topic outlier) and 30 topics for NMF. Therefore, the algebraic models would be the most suitable for modeling higher-level and broader topics on large text corpora to gain a general understanding of what the documents are about with

around 30–40 topics. Suitable datasets would be, for example, large collections of scientific papers, abstracts or articles because, according to the reviewed papers, short texts are not really suitable for this type of model.

The next type of model, which is the probabilistic LDA model, can be considered the middle ground between algebraic and LLM topic models regarding computational intensity. Nonetheless, the computational costs compared to algebraic models pay off by producing meaningful and interpretable topics. LDA performs well on various types of text data, but might face challenges with very short or rare-word-rich texts. This is a weakness that the model shares with LSA and NMF. If we also look at the average number of topics for LDA, we get around 30 topics. Given these findings, LDA would be best employed as an all-rounder for medium-sized document collections containing news, abstracts or extensive customer reviews to uncover more detailed topics in a reasonable amount of time.

Following, we have LLMs like the examined BERTtopic model, which delivers fine-grained topics that are the most interpretable in exchange for computational costs. It outperforms on noisy and sparse short-text data, where the other models struggle, but is limited to smaller datasets due to the intense computations. Therefore, we can conclude that the model would be best for getting granular insights from short social media posts, text messages or online reviews. The figure below illustrates the discovered relationship between computational costs and topic granularity.

Figure 4: Comparison of topic models



Source: Own results

We conclude our literature review and comparative analysis until now with the following overall findings: First, the development of the field of topic modeling was and is very much dependent on the evolution of data. This can be seen by examining the datasets used and the publication dates of the respective literature in our table, where we can spot a shift towards short-text data like tweets. Second, there is no one model that performs best on all tasks because every type of model has

advantages and disadvantages. The model choice depends on factors like corpus size, data characteristics or the type of topic to be modeled, e.g., broad or fine-grained. Evidence for this is provided by looking at the development of the field of topic modeling, the variety of models, and the performance of each approach regarding the different tasks.

What all models have in common is their reliance on and sensitivity to pre- or post-processing steps. LSA, NMF and LDA depend on a trial-and-error process that includes testing different preprocessing steps of the documents and also the selection of how many topics should be modeled. Also, the results from embedding-based models like BERTopic can differ according to the accompanying steps, like the choice of clustering algorithm in this case. Additionally, it is relatively hard to compare the different models and papers with each other given the various choices of parameters, evaluation metrics and data transformations, although some commonalities can be derived by examining the processing steps of each study. This will be further examined in the experimental section, but first we will move on to state-of-the-art LLMs.

2.7 State-of-the-Art LLMs

The emergence of GPT-3 initiated a paradigm shift in NLP by introducing an extraordinary capability to generate natural language texts that are astonishingly similar to those authored by humans. With 175B parameters and 96 layers trained on an extensive corpus comprising 499B tokens from web content, it surpassed its predecessor GPT-2 by more than a hundredfold in terms of size. Moreover, GPT-3 outperformed all other models available at that time in both the magnitude and coherence of the generated text. Notably, Microsoft's T-NLG and Google's T5-11B, which were the closest rivals during its launch, were merely a fraction of GPT-3's scale. (Dale, 2021)

A survey on GPT-3 found that the model is being applied in diverse domains such as developing conversational AI chatbots, software development, creative work and business productivity. Thereby, GPT-3 can generate product descriptions, advertisement headlines, blog ideas, email subject lines, and even poetic descriptions for images. Additionally, a potential application of GPT-3 in the healthcare domain, such as supporting customer service and triaging patients, is mentioned. (Zong & Krishnamachari, 2022)

2.7.1 Prompt Engineering

To interact with GPT-3, you typically provide it with a prompt or input text, and it generates a continuation or output text based on its learned patterns and structures of language. The quality and relevance of the output text can vary depending on the complexity and specificity of the prompt. Additionally, there are basically three methods for ”steering” an LLM like GPT-3 to perform the task at hand, namely zero-shot learning, few-shot learning, and fine-tuning.

Zero-shot learning does not involve any explicit training on a task; few-shot learning involves a small

amount of training data; and fine-tuning involves adapting the model to a specific task or domain using labeled data. Both zero-shot and few-shot learning are typically handled via a prompt (Brown et al., 2020) and basically during the inference step, whereas a fine-tuned model does not need a prompt after the additional training has happened. (OpenAI, 2023b)

Given the obvious advantages of not having to further train a model for a specific task, zero- and few-shot learning, and especially how to design prompts, are active fields of research. For example, a study done by Si and colleagues tried to improve the reliability of GPT-3 by developing simple and effective prompts that enhance its generalizability, social biases, calibration, and factuality. (Si et al., 2022) Another study done on ChatGPT, a model that is built on GPT-3 and fine-tuned for conversational interaction (OpenAI, 2022) developed a catalog of prompt engineering techniques in pattern form, providing reusable solutions to common problems. By identifying patterns in prompts, the catalog can help users design more effective prompts that are better suited for the task at hand. (White et al., 2023)

There is not only work done on examining general prompt engineering but also on finding domain-specific improvements. For example, a study done by Clavie and colleagues aimed to evaluate the impact of different aspects of prompt engineering on the performance of LLMs for job classification. The study found that prompt engineering, like optimizing the wording, is a critical factor in achieving high performance on the classification task and that the results are greatly influenced by elements of the prompt thought to be trivial. (Clavié et al., 2023) In another paper focusing on domain-specific prompt engineering, the author proposes the CLEAR framework, which should facilitate the interactions of students with ChatGPT for enhancing literacy education. (Lo, 2023) Another review on prompt engineering in healthcare further adds to the evidence that users can achieve significant improvements in accuracy and generate high-quality results by creating customized prompts that are designed for specific tasks and domains. (J. Wang et al., 2023)

To summarize these findings, we can conclude that the newest LLMs, especially GPT-3, enable a promising approach to a variety of NLP tasks. The possibility to leverage the pre-trained ability of these models to perform a variety of tasks with the help of carefully crafted prompts allows us to omit the tedious task of compiling labeled training datasets. Nonetheless, the optimal method of prompt engineering is currently an active field of research given the relatively recent emergence of GPT-3 and the fast iteration of the models, for example, ChatGPT and GPT-4. Given that the results of using a model like GPT-3 are very dependent on the prompt, a parallel can be drawn here to what we discovered regarding the robustness of the topic models that we discussed earlier on. Here also, the results are dependent on parameter selection or pre-processing, for example.

3 Experimental Section

The aim of the following experiment is to compare a novel ChatGPT-based topic modeling pipeline with two standard implementations using three different datasets. By evaluating the results quantitatively through measures such as topic coherence and topic diversity as well as qualitatively with the assistance of GPT-4, the objective is to determine whether the new pipeline outperforms the standard implementations on specific datasets.

To conduct this experiment, we will utilize three distinct datasets that cover news articles, scientific abstracts and tweets. Each dataset will be processed separately using both the novel ChatGPT-based pipeline and the two standard implementations. For quantitative evaluation, we will employ established measures such as topic coherence and topic diversity. Topic coherence will help us assess the degree of semantic consistency within each set of generated topics, while topic diversity will provide insights into the breadth and uniqueness of topics produced by each approach. The combination of these two metrics will allow us to get a better overall picture of the performance of the different topic modeling pipelines. For example, Dieng et al. defined the quality of topics as a combination of good coherence and diversity. (Dieng et al., 2020)

In addition to quantitative evaluation, we will leverage GPT-4, which is to date the most powerful model of the GPT family, to evaluate the quality of the generated topics. This qualitative assessment aims to capture subjective factors that may not be fully captured by quantitative measures alone. The usage of a state-of-the-art LLM for topic model evaluation was thereby first proposed by Rijcken et al. In their study, they showed that ChatGPT is able to create useful descriptions for topics generated by topic modeling algorithms. Thereby, they relied on the assessment of a domain expert, who deemed most of the outputs from ChatGPT to be at least somewhat useful. (Rijcken et al., 2023)

For our experiment, we will therefore rely on the more powerful GPT-4 to label the resulting topics and also to judge their quality. Through this experiment, we anticipate obtaining valuable insights regarding which implementation performs best across different datasets. The results will shed light on whether the novel ChatGPT-based pipeline surpasses standard implementations in terms of both quantitative measures and qualitative evaluations.

3.1 Datasets and Topic Models

For this experiment, we will be utilizing three different datasets representing documents that require general knowledge (news articles), expert knowledge (scientific abstracts), and also tweets, which are short-text data and a weak point of traditional topic models. Each dataset has been carefully selected based on the results of our literature review and to ensure variability and relevance to the topic modeling task.

1. **BBC Business News**: The first dataset is a subset of the BBC News dataset compiled by

Greene and Cunningham. (Greene & Cunningham, 2006) The business news dataset consists of 510 news articles, summing to a total word count of, 167729 and an average of 328 words per news article. The reason this dataset was chosen is that it provides a curated collection of news articles focusing solely on business-related articles, which facilitates topic interpretation. Although the choice of vocabulary is narrowed down by the business focus of the articles, the language in the documents is still comprehensible to a non-expert audience.

2. **Arxiv Scientific Abstracts**: Our second dataset represents documents that contain domain-specific vocabulary and require expert knowledge to comprehend. The dataset contains around 38972 entries (Sayak & Soumik, 2020) from which we sample 500 random abstracts that will be used for topic modeling. The 500 abstract subset contains, thereby, 85481 words in total and around 170 words on average per document. The choice for this dataset was again based on the literature review and to compare the topic models performances on documents with a more niche vocabulary.

3. **ChatGPT Tweets**: The last dataset that we will use for topic modeling consists of around 500 000 tweets about ChatGPT. (Ansari, 2023) This dataset is also reduced to a sample of around 1000 documents, which sums to a total of 25168 words and an average of around 25 words per tweet. This dataset choice is like all the others based on the results from the literature review and can be considered the ”stress test” for the traditional topic models and the proposed LLM-based pipeline.

According to the results compiled from the literature review, all three choices for the datasets are suitable for topic modeling and represent common document types in the field. Furthermore, although the size of the datasets is on the smaller side, they provide sufficient data points for conducting meaningful analysis while still being manageable in terms of computational requirements. The bottleneck thereby is the time associated with the inference of the ChatGPT model, which will be discussed in detail in the data preprocessing section.

The two models that will be our benchmark to compare the novel approach against are the LDA model, representing a probabilistic model, and the algebraic NMF model. Thereby, the selection of the models is based on the fact that they are widely used and their characteristics. Both models are suitable for a wide range of documents and have a weak point for short-text data. This opens up the possibility to compare the novel LLM-based pipeline to the performance of traditional models on established use cases and also to see if there is an improvement in modeling topics from current short-text data, which proves to be a challenge for standard approaches.

3.2 Pipeline Design

In this section, we will look at the design of the two different topic modeling pipelines that will be used in the comparison. First, we have a standard pipeline used with traditional topic models, and second, the novel LLM-based approach proposed in this thesis. We will discuss the implementation steps involved in both methods and explore how they uncover meaningful topics from the raw textual data.

First, we look at the overall steps that are required by both pipelines, which are the following:

1. **Data Preprocessing**: In this step, the raw text data is cleaned and preprocessed to remove any irrelevant or noisy information.
2. **Vectorization**: Once the data is preprocessed, it needs to be transformed into a numerical representation that can be used for modeling.
3. **Topic Modeling**: The vector representation of the documents is passed to the respective topic model algorithm, which results in a set of topics.
4. **Evaluation**: The resulting topics are evaluated with quantitative metrics and qualitative assessment.

3.2.1 Data Preprocessing

Arguably one of the most important parts is the data preprocessing or cleaning step. We can derive a suitable methodology for the standard pipeline by looking at the results of the literature review. Thereby, we can extract several important steps from the literature that we can employ in our own experiment. First, we have the simple lowercasing of all words and the removal of stop words, which are done to ensure that the text is clean and devoid of unnecessary noise. Removing stop words helps to focus on the more meaningful content of the text. Next, special and non-alphanumeric characters are removed to eliminate any unwanted symbols or punctuation marks that may interfere with the analysis process. These characters can include commas, periods, exclamation marks or symbols like the hashtag.

Following is the lemmatization of the remaining words, which plays a crucial role in simplifying and standardizing text for improved topic modeling. By reducing words to their base or dictionary form, this process enhances the accuracy of NLP techniques. Through lemmatization, the remaining words are transformed into a consistent format, enabling more effective data analysis. An example of a lemmatized word would be ”running” being transformed into its base form ”run”. An alternative to lemmatization would be stemming, but unlike lemmatization, which considers the part of speech and context to produce more meaningful results, stemming simply removes suffixes to derive the root form of a word. After lemmatizing, we also remove words with a character length smaller than 3, as these are mostly artifacts from lemmatizing or noise in the data that do not carry any significant meaning.

The last step in the process involves the formation of bi-grams. Bi-grams are pairs of consecutive words that help to capture the contextual meaning and relationships between words. The formation of bi-grams plays a crucial role in enhancing the comprehension of language. By capturing more context, bi-grams provide valuable insights into the relationships between words and improve the overall understanding of topics. An example of a bi-gram would be the combination of two consecutive words, such as ”natural language” or ”machine learning,” which can help extract deeper meaning from text and enable more accurate analysis.

After discussing the standard pipeline, we will now look at the preprocessing step for the LLM-based approach. The biggest difference is that the stop word removal, bi-gram and additional n-gram formation parts are done by the ChatGPT model in one step. This is achieved by prompting the LLM to extract keywords from the given document. Thereby, the LLM has the instruction to consider single or multiple words that contain a lot of context about the document as keywords. The formulation of the prompt follows best practice guidelines from OpenAI (Shieh, 2023) and can be found in the appendix.

This step is the most significant change in the novel topic modeling pipeline because the standard approach uses a fixed set of stop words and a bi-gram formation approach based on simple statistics. The idea here is to leverage the pre-training of the LLM to extract the words or phrases that carry the most meaning in the document. To fully leverage ChatGPT's capabilities and further improve the prompt, we also pass the type of document as context. For example, we clearly instruct the model that it should extract keywords from a business news article, not simply from a text. Additional steps that are done after keyword extraction are lowercasing and removal of words with two or fewer characters, for the same reason as in the standard preprocessing approach.

3.2.2 Vectorization

In order for machine learning algorithms to operate effectively, they require a numeric feature space. When dealing with text data, it becomes necessary to convert our documents into vectors for the purpose of performing machine learning tasks. This process is commonly known as feature extraction or simply vectorization. The Bag-of-Words (BoW) model stands out as the simplest type of document vectorizer, acknowledging that vocabulary contains both meaning and similarity. Though straightforward, this model is exceptional regarding its effectiveness and serves as the initial step towards more advanced models. When using a BoW strategy to vectorize a corpus, each document within the collection is transformed into a vector with dimensions matching the corpus's vocabulary size.

In order to produce these vectors, we make use of two alternative methods that we will use for both pipelines: term frequency (TF) vectorization and term frequency-inverse document frequency (TF-IDF) vectorization. Based on a review of the literature, it is evident that these two options are frequently utilized in topic modeling. The TF model takes a straightforward approach by populating the vector with the frequency of each word as it appears in the document. In this encoding method,

each document is represented as a multiset of its constituent tokens, and the count of each word position in the vector corresponds to its value.

Nevertheless, the TF approach solely considers a document independently without incorporating its context within the corpus. An alternative strategy would involve evaluating the relative frequency or scarcity of tokens in the document compared to their occurrence in other documents. The main idea is that meaning is most likely conveyed through the less common terms found within a document. The TF-IDF encoding method adjusts the token frequency in a document based on the frequency of the same token in the entire corpus. This technique emphasizes terms that are highly relevant to a particular instance. TF-IDF calculates the relevance of each token individually, taking into account its scaled frequency in the document and normalizing it by the inverse of its scaled frequency in the entire corpus. (Bengfort et al., 2018)

3.2.3 Topic Modeling

With the vectorized document collections, we can further progress to the actual topic modeling process. The standard pipeline will thereby employ both the LDA and the NMF models for generating topics. For the novel pipeline, we will take a similar approach to the word-embedding-based methodologies that we discussed earlier. Therefore, the topic modeling step consists of clustering the vectorized documents with the k-means algorithm. The choice for k-means was made because the algorithm allows you to choose the number of clusters, which represent topics in our case. This is very important and allows the two pipelines to be comparable in terms of topic coherence and diversity because we will use the number of topics as our variable parameter to optimize for the two metrics.

The k-means algorithm usually quickly and efficiently clusters a dataset within a few iterations and works as follows:

1. **Initialization**: The algorithm starts by randomly placing k centroids in the feature space, where k is the number of clusters you want to create.
2. **Assignment**: Each data point in the dataset is assigned to the cluster with the closest centroid. The distance between a data point and a centroid is typically calculated using Euclidean distance, but other distance metrics can also be used.
3. **Update**: After all data points have been assigned to clusters, the centroids are updated based on the mean of the data points assigned to each cluster. This means that the centroid is moved to the center of its assigned data points.

Steps 2 and 3 are repeated iteratively until the centroids no longer move significantly or a maximum number of iterations is reached. This ensures that the algorithm converges and the clusters become

stable. By repeating the assignment and update steps, the k-means algorithm aims to minimize the within-cluster sum of squares, also known as the inertia or distortion. This means that the algorithm tries to create clusters where the data points within each cluster are as close to each other as possible. Normally, the number of clusters would be chosen with the elbow method or the silhouette score (Géron, 2019), but for the use case of the experiment, we try to choose the number of clusters so that we maximize topic coherence and diversity to get higher-quality topics.

In the context of our topic modeling pipeline, the intuition behind the clustering approach is the following: Given that we created our vectors with either the TF or TF-IDF method, the clustering works by grouping the documents together based on their similarity or distance in the created feature space. The assumption is that documents that have similar TF or TF-IDF vectors are likely to contain similar topics. By retrieving the words that comprise the cluster centers, we can get a representation of what the documents in that cluster are about. In our case, we retrieve the top ten words from the cluster centers, which can be interpreted as representative keywords or themes for the documents in that cluster. They should give a sense of the main topics covered by the documents within that cluster.

Normally, the simple clustering of documents would not yield meaningful topics because the k-means algorithm was not developed with topic modeling in mind. Therefore, the approach relies on the preprocessing done with ChatGPT, which should be able to remove enough noise from the data to yield meaningful topics.

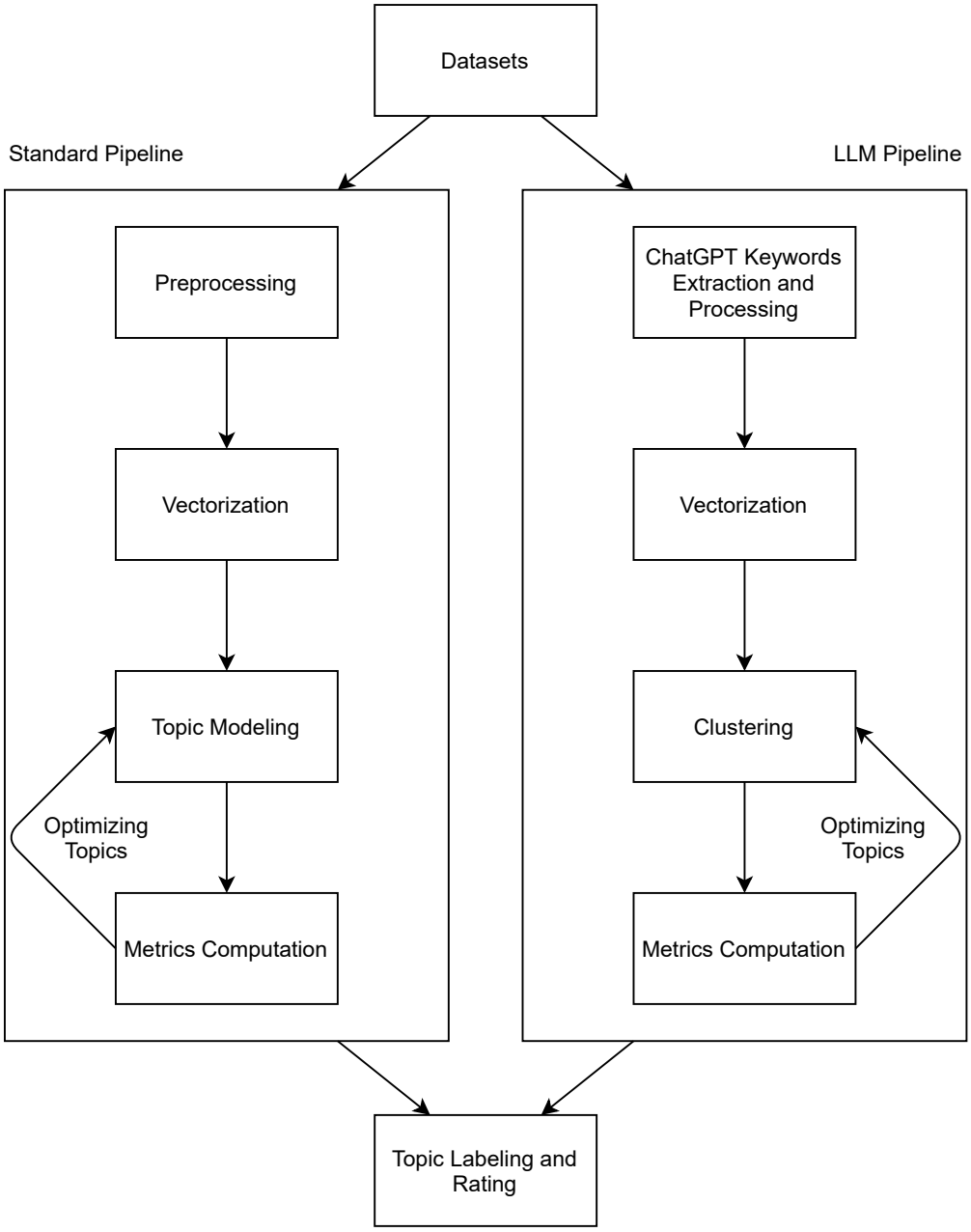
3.2.4 Evaluation

The final step of our pipeline is the evaluation of the resulting topics with quantitative and qualitative measures. To allow for choosing an optimum number of topics, both pipelines can be run for a range of topics. Thereby, the topic diversity and coherence measures of each run are plotted to examine their evolution over different numbers of topics. Additionally, we compute the best-performing number of topics for each parameter combination in terms of both metrics. The idea is then to further use the topics derived from the best-performing configuration for qualitative assessment with GPT-4, which takes on the role of our annotator.

The qualitative assessment is inspired by the approach used by Hoyle and colleagues in their study on automatic topic model evaluation. They asked human annotators to assess words from topic models regarding word intrusion and rate their coherence. (Hoyle et al., 2021) For our experiment, the actual instructions for the human annotators are repurposed as a prompt for the GPT-4 model. Thereby, we instruct GPT-4 to rate the relationship of the words as either very related, somewhat related or not very related, which is basically a qualitative coherence metric. In the second round of assessment, GPT-4 is tasked with counting the number of word intrusions for each list of topic words. The results are then averaged for every model-vectorizer combination. As a final step, we will then let GPT-4 label the topics. Before we move on to the implementation and results section, the following figure

summarizes the two topic modeling pipelines:

Figure 5: Topic modeling pipelines



Source: Own results

4 Implementation and Results

After discussing the topic modeling pipelines regarding their rationale, we will now look at the actual implementation and the results. Thereby, important technical details of the implementation, like libraries, algorithms and parameters, are discussed. Additionally, we will look at the results of the individual processing steps.

4.1 Corpus Preparation

As already mentioned, the initial datasets, especially the abstracts and tweets, were reduced to 500 and 1000 documents, respectively. This was done to match the business news dataset and also to reduce processing time and costs for the ChatGPT keyword extraction. The keyword extraction was thereby done with the OpenAI Python library (OpenAI, 2023a) which features functions for facilitating API calls towards their LLMs. For our purpose, we call the Chat Completions API of the GPT-3.5-Turbo model.

The respective prompt can be found in the appendix. At the time of writing, the costs for 1000 tokens are around €0,0014 for input and €0,0019 for output of the API. (OpenAI, 2023c) A token typically consists of around four characters, which would be around three-thirds of a standard English word. (OpenAI, 2023d) The following table summarizes the attributes, processing time and costs of the resulting keyword datasets.

Table 2: ChatGPT keywords extraction

Dataset	**Documents**	**Total Words**	**Average words per document**	**Remaining Words**	**Average keywords per document**	**Processing time (m)**	**API Costs (€)**
News	510	167729	328	31692	62	43:07	0.36
Abstracts	500	85481	170	26300	52	38:15	0.18
Tweets	1000	25168	25	10761	10	25:40	0.05

Source: Own results

It is important to note that the prompt does not instruct the model to extract a certain number of keywords but to extract the words and phrases that carry the most meaning. The words that are going to be extracted are based on what the model has learned regarding word importance during its initial training and the specific context of the document. This results in a greater reduction in document size for the larger documents with a lot of common words, like the news articles, and a smaller reduction for the smaller documents. Another critical factor to consider is the processing time, which is considerable. The processing time is thereby heavily dependent on the stability of the connection and the rate limit of the Chat Completions API. For the experiment, a two-second delay after every request was implemented to avoid reaching the rate limit imposed by OpenAI. Therefore, in an ideal circumstance without a rate limit, the actual processing times would be relatively faster. Next, we will look at the results of the preprocessing and vectorization steps, which are compiled in the next table below.

Table 3: Preprocessed datasets

Dataset	Pipeline	Total words	Average words per document	Vocabulary size
News	TTM	79430	155	8729
News	LLM	17432	34	10388
Abstracts	TTM	44479	88	5191
Abstracts	LLM	12801	25	9461
Tweets	TTM	9087	9	3029
Tweets	LLM	7332	7	4214

Source: Own results

Looking at Table 2 and Table 3, we can see that we get a significant reduction in total and average words after the cleaning step for both pipelines. The more significant difference between them is the ratio of total words to vocabulary size. Here, the LLM pipeline yields a much larger vocabulary because of the nature of the keyword extraction process and the fact that we did not lemmatize or stem the words. The keyword extraction can thereby generate n-grams that are naturally occurring in the documents and are not changed due to their aiding in the interpretability of the topics. A greater vocabulary size will result in more sparsity in the document vectors. If we look at the ratio of vocabulary size to total words for the tweets in the TTM pipeline, we can see that we have a ratio of around 1 to 3. Compared to the news, which has a ratio of around 1 to 10, and the abstracts, where we have a ratio of 1 to 9. According to the literature review, this predicts a worse performance for the tweets when using the TTM pipeline because both LDA and NMF are said to have a weakness regarding their ability to deal with sparse vectors.

Regarding the technical implementation, we use Python list comprehensions for tokenizing, lowercasing and removing words with special characters or numbers. The lemmatization is done with the NLTK library's WordNetLemmatizer, which uses the WordNet lexical database to identify the base form of a word. (NLTK Project, 2023). For bi-gram formation, we use the Phrases module from Gensim, which is a Python library for topic modeling. The Phrases library works by detecting and scoring co-occurrences of words in a given corpus. By analyzing the frequency and proximity of word pairs, it identifies bi-grams that occur more frequently than expected by chance. (Rehurek, 2023a)

4.2 Model Optimization

For the actual topic modeling, we use the Scikit-learn library's implementation of LDA, NMF and k-means. If we combine the possible combinations of datasets, vectorizers and topic models, we get eighteen different test series. We run each combination forty times on a range of 10 to 50 topics. The range was chosen with consideration of the results from Table 1. An important note is that all other parameters of the respective algorithms were left at their standard values because tuning all hyperparameters is outside the scope of this experiment. The same is true for the vectorization and preprocessing steps where external libraries were used.

The best-performing number of topics for the respective vectorizer-model combination was chosen based on a ranking method. Thereby, two lists are created, representing either coherence or diversity, with each list being sorted in descending order. The number of topics that ranks the highest for both metrics is considered the winner. Table 4 showcases the results of the topic modeling runs, with the top performers for each dataset written in bold font.

Table 4: Quantitative topic modeling performance

Dataset	Vectorizer	Model	Best Performing Topics	Coherence	Diversity	Time (s)
News	TF	lda	12	-21.14	0.54	82.8
News	TF	nmf	10	-20.95	0.76	55.5
News	TF	kmeans	47	-21.41	0.93	17.3
News	**TF-IDF**	**lda**	**17**	**-20.65**	**0.96**	**38.1**
News	TF-IDF	nmf	15	-20.63	0.91	55.4
News	TF-IDF	kmeans	25	-21.24	0.89	14.6
Abstracts	TF	lda	11	-20.72	0.43	53.3
Abstracts	TF	nmf	19	-20.58	0.75	31.3
Abstracts	TF	kmeans	20	-21.13	0.96	14.2
Abstracts	**TF-IDF**	**lda**	**14**	**-20.38**	**0.93**	**27.5**
Abstracts	TF-IDF	nmf	11	-20.64	0.87	35.4
Abstracts	TF-IDF	kmeans	42	-21.4	0.89	13.9
Tweets	TF	lda	10	-20.87	0.63	37.7
Tweets	TF	nmf	10	-20.75	0.82	16.0
Tweets	**TF**	**kmeans**	**30**	**-20.69**	**0.83**	**13.4**
Tweets	TF-IDF	lda	37	-20.93	0.76	28.4
Tweets	TF-IDF	nmf	14	-20.98	0.9	22.7
Tweets	**TF-IDF**	**kmeans**	**19**	**-20.62**	**0.82**	**14.6**

Source: Own results

Topic coherence is computed with the help of the Gensim library's implementation of the UMass metric. (Rehurek, 2023b) The UMass metric was chosen because of its frequent appearance in the literature and its fast computation. Regarding the interpretation of the metric, a higher score indicates more coherent topics, with the scores being negative in general. (Thielen, 2022) The computation of the topic diversity metric is adapted from Dieng et al. with the change of using only the top ten words instead of 25. A lower diversity is thereby considered inferior to a diversity close to one. (Dieng et al., 2020) The figures displaying the evolution of the metrics for different numbers of topics can be found in the appendix.

4.3 Topic Interpretation

Like already explained, the interpretation and qualitative assessment of the topics were done with the more powerful GPT-4 model. The model was prompted via the Chat Completions API to assign a coherence rating, detect word intrusions and label the topics. The respective prompts and the labeled topics can be found in the appendix. Table 5 showcases the average ratings from this assessment.

Like in Table 4, the top performers are marked in bold.

Table 5: Qualitative topic ratings

Dataset	Vectorizer	Model	Rating	Intrusions
News	tf	lda	1.58	1.42
News	tf	nmf	1.1	1.0
News	tf	kmeans	1.34	1.6
News	tfidf	lda	2.06	2.76
News	**tfidf**	**nmf**	**1.07**	**0.73**
News	tfidf	kmeans	1.16	1.24
Abstracts	tf	lda	1.09	0.91
Abstracts	tf	nmf	1.16	1.0
Abstracts	tf	kmeans	1.3	1.9
Abstracts	tfidf	lda	2.0	2.71
Abstracts	**tfidf**	**nmf**	**1.0**	**0.82**
Abstracts	tfidf	kmeans	1.17	1.31
Tweets	**tf**	**lda**	**1.9**	**1.8**
Tweets	tf	nmf	1.9	1.9
Tweets	**tf**	**kmeans**	**1.57**	**2.0**
Tweets	tfidf	lda	2.19	2.3
Tweets	tfidf	nmf	2.0	2.0
Tweets	tfidf	kmeans	1.68	2.21

Source: Own results

4.4 Discussion

Overall, the results of the experiment can be considered to follow the trend of what could be observed in the field of topic modeling as a whole. There is no universal approach for measuring performance, and different measuring criteria tend to disagree with each other. Although we can definitely observe a trend when we compare the results from the quantitative and qualitative analyses. Here the traditional algorithms with TF-IDF weighting seem to perform better on the news and abstracts, whereas TF vectorizing and our novel LLM-based approach outperform on the tweets. The only outlier that we have is the TF-LDA combination for the tweets in the qualitative assessment, which subjectively underperforms our novel approach. To illustrate this, we can look at a topic sample of either method.

Table 6: Tweets TF-LDA Topics Sample

Topics	Labels
data chatgpt human think tool answer say tell anything going	Artificial Intelligence Communication
chatgpt model language make language_model thought game new via like	ChatGPT and Language Modeling
chatgpt used application great research way buy based potential trial	AI Technology Use & Potential
chat gpt chat_gpt like ask using get write make use	Chatbot Functions
using read use much artificial artificial_intelligence intelligence take think make	Artificial Intelligence Usage

Source: Own results

Table 7: Tweets TF-k-means Topics Sample

Topics	Labels
websites apps launch discord smart_tech implement user_experience bandwagon chatgpt enterprises	Digital Technology and User Experience
chatgpt openai artificial_intelligence google microsoft gpt chatbot technology future bing	Artificial Intelligence and Technology Companies
chat_gpt technical_seo_tips higher google video search_engine_optimization online_marketing seo wordpress_seo_tutorial google_ranking	Digital Marketing and SEO
spit_out entire_internet seconds prompt recreate learning_language_patterns chatbot coherent chatgpt collective_works	Artificial Intelligence and Language Processing
nft mysticism poem openai haiku writing evolve writer attempt logic	Artificial Intelligence and Creativity

Source: Own results

By looking at these two samples, we can derive several conclusions. First, the labeling is difficult for the outputs of both models. If there is a topic that is also clear to label for a human, then GPT-4 also labels it clearly most often. For example, if we look at topic three in Table 7, we can see that the topic has both good interpretability and labeling. In general, the topics derived from the novel pipeline can be better interpreted, for example, topic 1 could be about the usage of chatbots to enhance the user experience, and topic 2 could be about competition for search engines between large technology corporations. On the other hand, topics 4 and 5 talk about ChatGPT capabilities regarding information retrieval and writing.

Looking at the topics of the TF-LDA combination, it is much harder to gain any deeper insights into the dataset. For example, Topic 1 of the TF-LDA combination could be about speculations about ChatGPT thinking like a human, and Topic 2 could be about its usage in research. Of course, these are only subjective opinions with a vague interpretation. Nonetheless, the results of the quantitative and qualitative analyses seem to agree. Additionally, if we read over the topics that can be found in the appendix, we can come to the same conclusion.

Of course, there are a lot of variables that could be changed, which would probably have a significant influence on the results. For example, choosing other quantitative metrics or adding additional ones, optimizing additional parameters of the respective algorithms, or using alternative algorithms altogether. Another improvement could possibly be achieved by combining keywords with similar meaning but different wording into a single keyword. This could be done, for example, by creating word embeddings and computing their cosine similarity. With this additional step, we could effectively reduce the vocabulary in the same way as lemmatization, which would probably yield an even better quality of the topics for the LLM pipeline. Also, the prompts for either keyword extraction or topic labeling and assessment could be worded differently, which seems to have a significant effect on the output. There would also be the possibility to fine-tune a model for either of the two tasks, which would probably result in a better performance. Additionally, we could run the qualitative assessment

a number of times with the same approach and compile the results to get an average.

As an alternative to the qualitative examination with a LLM, we could also employ a survey where humans are tasked with labeling the topic words and rating their interpretability. According to the literature review, this would be considered the gold standard, but regarding implementation, this approach would also have some drawbacks. For example, a survey is most often associated with the time of human annotators, who may or may not be experts in the respective domain of the data that we try to model. Therefore, it would be of great value to develop a qualitative rating approach based on a LLM. This could be done by either fine-tuning a model for this task or, for example, running a similar setup like in our experiment a number of times. An interesting approach going in this direction would be to inject even more randomness into the prompt by assigning different knowledge backgrounds to the annotators we try to simulate with the LLM. This would be close to the circumstances of a real human survey.

Coming back to the initial thought of using a LLM for evaluating topic modeling results, the experiment and other studies already showed that, at least in its basic form, this approach would be a possibility, although the reliability of the ratings is only somewhat useful with the configuration employed in this experiment. Nonetheless, the new state-of-the-art models from GPT-4 onward definitely have the capabilities for such a task.

A different aspect and probably more of a bottleneck for now is the inference time and costs associated with the novel topic modeling pipeline, which would be a considerable obstacle for larger datasets. If we extrapolate the values from Table 2 to, for example, a dataset with a million tweets, we would have a processing time of around seventeen days. Although, the question that arises is whether analyzing such large datasets is even the purpose of the novel pipeline. Like we already discussed in the previous sections, there seems to be a trade-off between interpretability and computational costs between the traditional and LLM-based topic modeling approaches. Interestingly enough, the quantitative analysis favors a larger number of topics for the novel pipeline in general, as can be seen in Table 4. This further adds to the evidence that a LLM-based approach is probably better suited for deriving a higher number of fine-grained topics from a smaller dataset. In contrast, traditional approaches can complement this by giving a higher-level overview of larger datasets.

5 Conclusion

In conclusion, this research aimed to address three main questions regarding the distinctive characteristics of LLMs and traditional statistical algorithms when employed for topic modeling. Their performance in terms of computational intensity and topic quality, and the implications of choosing one approach over the other based on specific goals, dataset characteristics, and available resources.

From the literature review and experimental analysis conducted in this study, it is evident that tradi-

tional topic models excel at handling larger document collections with longer texts, such as news or research papers. However, they prove to be less effective in dealing with new short-text datasets, which are often noisy. On the other hand, LLM-based approaches demonstrate better capability for deriving useful topics from noisy datasets, albeit requiring more computational resources. The quantitative and qualitative analyses support these findings by favoring traditional topic modeling algorithms for news and abstracts, while highlighting the novel LLM-based approach proposed in the thesis as more suitable for tweets. It is important to note that the current LLM-based pipeline based on GPT models is not yet feasible for large datasets and that it is difficult to derive general conclusions due to the large number of variables that could be changed in the experiment.

Looking ahead, as LLMs continue to become more efficient with better training data, for example, as Gunasekar et al. showed in their study on the impact of data quality on model performance and training costs (Gunasekar et al., 2023), in the future they may become a viable option even for larger datasets. Hence, the preliminary recommendation for now is to utilize traditional models for larger datasets to achieve a higher-level overview and employ the new pipeline using the GPT model family for obtaining more granular insights into smaller document collections. In summary, this study recognizes the strengths and limitations of both LLM-based approaches and traditional statistical algorithms in different contexts. The choice between these approaches should be carefully considered based on specific goals, dataset characteristics, and resource availability. By utilizing both methods strategically, researchers can obtain comprehensive topic modeling results across various types of documents.

Finally, it must be acknowledged that further research is needed to explore the potential applications and optimize the performance of LLMs in order to fully harness their capabilities. Additionally, the limitations and challenges identified throughout this research process should serve as valuable insights for future studies in the field of topic modeling.

References

Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., & Hassan, A. (2023). Topic modeling algorithms and applications: A survey. *Information Systems*, *112*, 102131.

Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in artificial intelligence*, *3*, 42.

Alghamdi, R., & Alfalqi, K. (2015). A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications*, *6*(1), 147–153.

Ansari, K. (2023). *500k chatgpt-related tweets jan-mar 2023.* Retrieved from `https://www.kaggle.com/datasets/khalidryder777/500k-chatgpt-tweets-jan-mar-2023`

Athukorala, S., & Mohotti, W. (2022). An effective short-text topic modelling with neighbourhood assistance-driven nmf in twitter. *Social Network Analysis and Mining*, *12*(1), 89.

Azzopardi, L., Girolami, M., & van Risjbergen, K. (2003). Investigating the relationship between language model perplexity and ir precision-recall measures. In *Proceedings of the 26th annual international acm sigir conference on research and development in informaion retrieval* (pp. 369–370).

Bellaour, S., Bellaour, M. M., & Ghada, I. E. (2021). Topic modeling: Comparison of lsa and lda on scientific publications. In *4th international conference on data storage and data engineering* (pp. 59–64).

Bengfort, B., Bilbro, R., & Ojeda, T. (2018). *Applied text analysis with python: Enabling language-aware data products with machine learning*. Sebastopol: O'Reilly Media.

Bergamaschi, S., & Po, L. (2015). Comparing lda and lsa topic models for content-based movie recommendation systems. In *International conference on web information systems and technologies* (pp. 247–263).

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Blei, D. M., & Lafferty, J. D. (2005). Correlated topic models. In *Proceedings of the 18th international conference on neural information processing systems* (pp. 147–154).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*, 993–1022.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. In *Proceedings of the 34th international conference on neural information processing systems* (pp. 1877–1901).

Campagnolo, J. M., Duarte, D., & Dal Bianco, G. (2022). Topic coherence metrics: How sensitive are they? *Journal of Information and Data Management*, *13*(4).

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. In *Proceedings of the 22nd international conference on neural*

information processing systems (pp. 288–296).

Chen, Y., Zhang, H., Liu, R., Ye, Z., & Lin, J. (2019). Experimental explorations on short text topic mining between lda and nmf based schemes. *Knowledge-Based Systems*, *163*, 1–13.

Chuang, J., Gupta, S., Manning, C. D., & Heer, J. (2013). Topic model diagnostics: assessing domain relevance via topical alignment. In *Proceedings of the 30th international conference on international conference on machine learning* (Vol. 28, pp. 612–620).

Churchill, R., & Singh, L. (2022). The evolution of topic modeling. *ACM Computing Surveys*, *54*(10s), 1–35.

Clavié, B., Ciceu, A., Naylor, F., Soulié, G., & Brightwell, T. (2023). *Large language models in the workplace: A case study on prompt engineering for job type classification.* Retrieved from `https://arxiv.org/abs/2303.07142`

Crossno, P., Dunlavy, D., & Shead, T. (2009). Lsaview: A tool for visual exploration of latent semantic modeling. In *Ieee symposium on visual analytics science and technology* (pp. 83–90).

Cvitanic, T., Lee, B., Song, H. I., Fu, K., & Rosen, D. (2016). Lda v. lsa: A comparison of two computational text analysis tools for the functional categorization of patents. In *International conference on case-based reasoning* (pp. 41–50).

Dale, R. (2021). Gpt-3: What's it good for? *Natural Language Engineering*, *27*(1), 113–118.

David, N., Jey Han, L., Kar, I. G., & Timothy, B. (2010). Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics* (pp. 100–108).

de Groot, M., Aliannejadi, M., & Haas, M. R. (2022). *Experiments on generalizability of bertopic on multi-domain short text.* Retrieved from `https://arxiv.org/abs/2212.08459`

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Delvin, C. Z., & Hady, L. (2022). Dynamic topic models for temporal document networks. In *Proceedings of the 39th international conference on machine learning* (pp. 26281–26292).

de Waal, A., & Barnard, E. (2008). Evaluating topic models with stability. In *Nineteenth annual symposium of the pattern recognition association of south africa* (pp. 79–84).

Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 439–453.

Du, L., Buntine, W., Jin, H., & Chen, C. (2012). Sequential latent dirichlet allocation. *Knowledge and Information Systems*, *31*(3), 475–503.

Egger, R., & Yu, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, *7*, 886498.

Evangelopoulos, N., Zhang, X., & Prybutok, V. R. (2012). Latent semantic analysis: five methodolog-

ical recommendations. *European Journal of Information Systems*, *21*, 70–86.

Foltz, P. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, & Computers*, *28*, 197–202.

Géron, A. (2019). *Hands-on machine learning with scikit-learn, keras, and tensorflow: Concepts, tools, and techniques to build intelligent systems* (2nd ed.). Sebastopol: O'Reilly Media.

Greene, D., & Cunningham, P. (2006). Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd international conference on machine learning* (pp. 377–384).

Greene, D., O'Callaghan, D., & Cunningham, P. (2014). How many topics? stability analysis for topic models. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 498–513).

Grootendorst, M. (2022). *Bertopic: Neural topic modeling with a class-based tf-idf procedure.* Retrieved from `https://arxiv.org/abs/2203.05794`

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Del Giorno, A., Gopi, S., ... Li, Y. (2023). *Textbooks are all you need.* Retrieved from `https://arxiv.org/abs/2306.11644`

Haddock, J., Kassab, L., Li, S., Kryshchenko, A., Grotheer, R., Sizikova, E., ... Leonard, K. (2020). *Semi-supervised nmf models for topic modeling in learning tasks.* Retrieved from `https://arxiv.org/abs/2010.07956`

Hanna M., W. (2008). *Structured topic models for language*. University of Cambridge.

Hanna M. Wallach. (2006). Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on machine learning* (pp. 977–984).

Hofmann, T. (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).

Hong, L., Yin, D., Guo, J., & Davison, B. D. (2011). Tracking trends: incorporating term volume into temporal topic models. In *Proceedings of the 17th acm sigkdd international conference on knowledge discovery and data mining* (pp. 484–492).

Hoyle, A., Goel, P., Peskov, D., Hian-Cheong, A., Boyd-Graber, J., & Resnik, P. (2021). Is automated topic model evaluation broken?: The incoherence of coherence. In *Advances in neural information processing systems 34* (pp. 2018–2033).

Kaur, A., & Kumar, M. (2019). Performance analysis of lsa for descriptive answer assessment. In *Innovations in computer science and engineering* (Vol. 74, pp. 57–63).

Kherwa, P., & Bansal, P. (2019). Topic modeling: A comprehensive review. *EAI Endorsed Transactions on Scalable Information Systems*, *7*(24), 159623.

Korenčić, D., Ristov, S., Repar, J., & Šnajder, J. (2021). A topic coverage approach to evaluation of topic models. *IEEE Access*, *9*, 123280–123312.

Kuang, D., Choo, J., & Park, H. (2015). Nonnegative matrix factorization for interactive topic modeling and document clustering. *Partitional clustering algorithms*, 215–243.

Li, D., Zhang, J., & Li, P. (2019). Tmsa: A mutual learning model for topic discovery and word embedding. In *Proceedings of the 2019 siam international conference on data mining* (pp. 684–692).

Lidan, Z. (2022). Topic modeling based on attributed graph. In *2022 19th international computer conference on wavelet active media technology and information processing* (pp. 1–4).

Likhitha, S., Harish, B. S., & & Kumar, H. K. (2019). A detailed survey on topic modeling for document and short text data. *International Journal of Computer Applications*, *178*(39), 1–9.

Lo, L. S. (2023). The clear path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship*, *49*(4), 102720.

Matthews, P. (2019). Human-in-the-loop topic modelling: Assessing topic labelling and genre-topic relations with a movie plot summary corpus. In *The human position in an artificial world: Creativity, ethics and ai in knowledge organization* (pp. 181–207).

Mcauliffe, J., & Blei, D. M. (2007). Supervised topic models. In *Advances in neural information processing systems 20* (pp. 121–128).

Mohammed, S., & Al-augby, S. (2020). Lsa & lda topic modeling classification: Comparison study on e-books. *Indonesian Journal of Electrical Engineering and Computer Science*, *19*(1), 2502–4752.

Naili, M., Habacha, A., & Ben Ghezala, H. (2018). Parameters driving effectiveness of lsa on topic segmentation. In *Computational linguistics and intelligent text processing: 17th international conference* (pp. 560–572).

Nikolenko, S. I. (2016). Topic quality metrics based on distributed word representations. In *Proceedings of the 39th international acm sigir conference on research and development in information retrieval* (pp. 1029–1032).

NLTK Project. (2023). *Wordnetlemmatizer.* Retrieved from `https://www.nltk.org/api/nltk.stem.wordnet.html#nltk.stem.WordNetLemmatizer`

O'Callaghan, D., Greene, D., Carthy, J., & Cunningham, P. (2015). An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, *42*(13), 5645–5657.

OpenAI. (2022). *Chatgpt.* Retrieved from `https://openai.com/blog/chatgpt`

OpenAI. (2023a). *Chat completions api.* Retrieved from `https://platform.openai.com/docs/guides/gpt/chat-completions-api`

OpenAI. (2023b). *Fine-tuning.* Retrieved from `https://platform.openai.com/docs/guides/fine-tuning`

OpenAI. (2023c). *Pricing.* Retrieved from `https://openai.com/pricing`

OpenAI. (2023d). *Tokenizer.* Retrieved from `https://platform.openai.com/tokenizer`

Pandey, R., & Mohler, G. (2018). Evaluation of crime topic models: topic coherence vs spatial crime concentration. In *2018 ieee international conference on intelligence and security informatics* (pp. 76–78).

Papadia, G., Pacella, M., Perrone, M., & Giliberti, V. (2023). A comparison of different topic modeling methods through a real case study of italian customer care. *Algorithms*, *16*(2), 94.

Qiang, J., Qian, Z., Li, Y., Yuan, Y., & Wu, X. (2022). Short text topic modeling techniques, applications, and performance: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *34*(3), 1427–1445.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2018). *Language models are unsupervised multitask learners.* Retrieved from https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Rehurek, R. (2023a). *Phrases.* Retrieved from https://radimrehurek.com/gensim/models/phrases.html

Rehurek, R. (2023b). *Umass.* Retrieved from https://github.com/RaRe-Technologies/gensim/blob/develop/gensim/topic_coherence/direct_confirmation_measure.py#L19

Rijcken, E., Scheepers, F., Zervanou, K., Spruit, M., Mosteiro, P., & Kaymak, U. (2023). Towards interpreting topic models with chatgpt. In *The 20th world congress of the international fuzzy systems association.*

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth acm international conference on web search and data mining* (pp. 399–408).

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2012). *The author-topic model for authors and documents.* Retrieved from https://arxiv.org/abs/1207.4169

Rüdiger, M., Antons, D., Joshi, A. M., & Salge, T.-O. (2022). Topic modeling revisited: New evidence on algorithm performance and quality metrics. *PloS one*, *17*(4), e0266325.

S. Suh, J. Choo, J. Lee, & C. K. Reddy. (2016). L-ensnmf: Boosted local topic discovery via ensemble of nonnegative matrix factorization. In *2016 ieee 16th international conference on data mining* (pp. 479–488).

Sayak, P., & Soumik, R. (2020). *Large-scale multi-label text classification.* Retrieved from https://keras.io/examples/nlp/multi_label_classification/

Shieh, J. (2023). *Best practices for prompt engineering with openai api.* Retrieved 29.09.2023, from https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api

Si, C., Gan, Z., Yang, Z., Wang, S., Wang, J., Boyd-Graber, J., & Wang, L. (2022). *Prompting gpt-3 to be reliable.* Retrieved from https://arxiv.org/abs/2210.09150

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*,

427(7), 424–440.

Su, J., Boydell, O., Greene, D., & Lynch, G. (2015). *Topic stability over noisy sources.* Retrieved from https://arxiv.org/abs/1508.01067

Suh, S., Shin, S., Lee, J., Reddy, C. K., & Choo, J. (2018). Localized user-driven topic discovery via boosted ensemble of nonnegative matrix factorization. *Knowledge and Information Systems*, *56*(3), 503–531.

Suri, P., & Roy, N. (2017). Comparison between lda & nmf for event-detection from large text stream data. In *2017 3rd international conference on computational intelligence & communication technology* (pp. 1–5).

Teh, Y., Jordan, M., Beal, M., & Blei, D. (2004). Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in neural information processing systems 17* (pp. 1385–1392).

Thielen, B. (2022). *Generating topic models from corpora across languages*. University of Liege.

Thompson, L., & Mimno, D. (2020). *Topic modeling with contextualized word representation clusters.* Retrieved from https://arxiv.org/abs/2010.12626

Tijare, P., & Rani, P. (2020). Exploring popular topic models. In *Journal of physics: Conference series* (Vol. 1706, p. 012171).

Wang, J., Shi, E., Yu, S., Wu, Z., Ma, C., Dai, H., ... Zhang, S. (2023). *Prompt engineering for healthcare: Methodologies and applications.* Retrieved from https://arxiv.org/pdf/2304.14670

Wang, Y., Liu, J., Qu, J., Huang, Y., Chen, J., & Feng, X. (2014). Hashtag graph based topic model for tweet mining. In *2014 ieee international conference on data mining* (pp. 1025–1030).

Wang, Z., Li, L., Zhang, C., & Huang, Q. (2015). Image-regulated graph topic model for cross-media topic detection. In *Proceedings of the 7th international conference on internet multimedia computing and service* (pp. 1–4).

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., ... Schmidt, D. C. (2023). *A prompt pattern catalog to enhance prompt engineering with chatgpt.* Retrieved from https://arxiv.org/pdf/2302.11382

Xie, P., Yang, D., & Xing, E. (2015). Incorporating word correlation knowledge into topic modeling. In *Proceedings of the 2015 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 725–734).

Xu, M., Yang, R., Ranshous, S., Li, S., & Samatova, N. F. (2017). Leveraging external knowledge for phrase-based topic modeling. In *2017 conference on technologies and applications of artificial intelligence* (pp. 29–32).

Xu, X., Stulp, G., van den Bosch, A., & Gauthier, A. (2022). Understanding narratives from demographic survey data: a comparative study with multiple neural topic models. In *Proceedings of the fifth workshop on natural language processing and computational social science* (pp.

33–38).

Yang, H., Xinhuai, T., Tiancheng, T., Yunlong, H., & Jintai, T. (2020). Enhancing topic models by incorporating explicit and implicit external knowledge. In *Asian conference on machine learning* (pp. 353–368).

Zengul, F., Bulut, A., Oner, N., Ahmed, A., Yadav, M., Gray, H. G., & Ozaydin, B. (2023). A practical and empirical comparison of three topic modeling methods using a covid-19 corpus: Lsa, lda, and top2vec. In *Proceedings of the 56th hawaii international conference on system sciences* (pp. 930–939).

Zhao, H., Phung, D., Huynh, V., Jin, Y., Du Lan, & Buntine, W. (2021). *Topic modelling meets deep neural networks: A survey.* Retrieved from `https://arxiv.org/abs/2103.00498`

Zhao, X., Wang, D., Zhao, Z., Liu, W., Lu, C., & Zhuang, F. (2021). A neural topic model with word vectors and entity vectors for short texts. *Information Processing & Management*, *58*(2), 102455.

Zong, M., & Krishnamachari, B. (2022). *A survey on gpt-3.* Retrieved from `https://arxiv.org/abs/2212.00857`

Appendices

Appendix directory

Appendix A: Topics

BBC_News-tf-lda

Topics	Labels
car said sale profit cost year sri people new market	Economy and Automobile Industry
said would deutsche offer boerse bid lse deutsche_boerse takeover shareholder	Business Acquisition
said work million last company people also year borussia jobless	Business and Economy
said price sale oil new also year club crude united	Business and Economy
said rate growth economy economic price bank figure year rise	Economy and Finance
said airline new would company european last could also cost	Business and Aviation
yukos said russian oil company court firm tax bank sale	Russian Business and Legal Issues
said fiat company firm ebbers former worldcom also fraud business	Corporate Fraud
said firm company also new would share country year government	Business and Government Relations
share said firm market stock year new company marsh card	Business and Finance
said bank firm former china government company also new two	Business and Economy
said project government company firm one new state investment told	Business and Government Affairs

BBC_News-tf-nmf

said would company analyst debt country also people minister group	Economics and Politics
rate growth economy interest said interest_rate consumer economist economic job	Economics
yukos russian oil court company tax bankruptcy gazprom auction firm	Russian Business and Law
sale profit year car store retail rise said last rose	Business and Economy
bank standard china south chartered banking standard_chartered foreign financial one	Banking and Finance
price oil house house_price said crude fall market rise mortgage	Economic Trends in Housing and Oil Market
deutsche boerse deutsche_boerse lse bid offer shareholder would euronext stock	Stock Market and Trading
share market stock firm company stock_market investor profit analyst financial	Finance and Stock Market
new firm last could airline executive company said chief year	Business and Corporate News
economic government budget economy also year deficit tax spending state	Economics and Government Finance

BBC_News-tf-kmeans

Topics	Labels
consumer_prices peg government_efforts renminbi_revaluation producer_prices lending exports investment competitiveness domestic_demand	Economic Factors and Policies
annual_turnover roaming_market t_mobile allegations mobile_call_charges excessive uk_roaming_rates dominant_market_position deutsche_telekom foreign_mobile_operators	Mobile Communication Industry
figures northern_ireland rise report halifax mortgage_lender scotland british_property family_home tim_crawford	UK Real Estate Market
telecoms_boom technology_firms com_survivors otc_bulletin_board stock_index amazon tech_giants registration_document raise nasdaq	Tech Industry and Stock Market

Topics	Labels
hold housing_market_outlook corrective_action british_chambers_of_commerce british_retail_consortium house_prices economy manufacturers manufacturing eef	UK Economy and Housing Market
euronext deutsche_boerse shareholders bid lse offer cash talks transaction_fees headquarters	Stock Exchange Acquisition
losses profitability profits annual_savings japanese_subsidiary stake luxury_car_maker s_most_successful_luxury_brand american_express analyst_expectations	Business and Finance
italy bankruptcy_laws industrialists competitiveness_of_small_firms trade_union_leaders corporate_landscape entrepreneurial_spirit _spending government_spending spending	Italian Economy
economic_growth budget poor health distribution subsidies government_expenditure education disadvantaged productive_investment	Government Budget and Economy
real_experience chief_operating_officer space_invaders news_corp takeover_list feature_films peter_chernin pac_man development_companies profitable	Video Games and Film Industry
baikal yuganskneftegas sale auction collapse sanctions oil_producer ownership worldwide_jurisdiction s_national_petroleum_corporation	Oil Industry and Trade
saudi_arabia riyadh damage_cover saudi_ncci listing ncci government_warning jeddah boost audited_accounts	Saudi Arabia Business and Finance
lighting distribution_capacity congeniality walls water_supply wall_lights wall_insulation uplights offices office_work	Office Infrastructure
toyota ford s_top_car_maker s_daimlerchrysler calendar_year foreign_rivals petrol_price_conscious_consumers production sales_increase asian_assault	Automobile Industry
analyst new_york_court corporate_wrongdoing bizarre predicted terrorist_attacks witness_tampering negative_publicity crime_fighting_tools sentences	Legal Issues and Crime in Corporate America
profits creating global_recorded_music_industry currency coldplay music_sales physical_music_market prior_year disappointing_sales digital_music	Music Industry Economics
shares chief_executive analysts growth china exports economy sales economic_growth government	Business and Economy
sri_lanka india thailand real_estate_sectors developing_countries world_bank tragedy jobs south_east_asian_region recovery	South East Asian Economy and Development
bdo_stoy_hayward s_output_index researchers inflation_index bdo uncertainties peter_hemington interest_rate_rises optimism merger_and_acquisition_activity	Economics and Business Research
opponent soviet_union rock_bottom_prices investors state_hands jilted_bidders foreign_groups stuart_hensel foreign_investment economist_intelligence_unit	Economic Policies and Foreign Investment
foreign_exchange_rates building_trade jobs nano_technology domestic_demand redundancies oil_prices credit_bubble mortgage_loans bankruptcy	Economy and Finance
ongc buy yukos_assets chinese_crude_company oil_production s_levels managing_director stagnated battling moscow	Oil Industry and Business
dollar euro interest_rates us_economy consumer_spending federal_reserve budget_deficits yen economic_growth oil_prices	Economics and Finance

Topics	Labels
rose imports s_economic_policies us_trade_gap oil china president_bush marie_pierre_ripert faster_rate increasing_domestic_demand	US-China Trade Relations
military_operations worse_than_expected dollar_investors administration fiscal_year additional_funds improvement us_exports us_economic_growth twin_deficit	US Economy and Trade
patience it_firm call_centres call_centre_workers call_centre_users risk current_accounts _society welcome_message brand_damage	Customer Service in IT Sector
china iraq political_leaders iran corporate_leadership world_leaders world_social_forum developing_countries vaccination_campaign deep_frost	International Politics and Leadership
opposition open_letter lee_scott urban_legends research_data faulty_data products chief_executive average_pay fda	Business and Research Controversies
worldcom verizon buyout highest_level future_growth_prospects bid us_phone_industry successful_mobile_division chief_executive billion_dollar_telecoms_deal	Telecommunications Industry
russian_government mikhail_khodorkovsky state_control recover forced_sale repay warned the_hague accused offshore_firms	Russian Politics and Business
takeover oatleys rosemount_estates brl_hardy price rival southcorp robert_mondavi beer_brands board	Wine and Beer Business
shareholders fraud yukos russia rosneft bankruptcy_protection auction chief_executive gazprom trial	Russian Business and Legal Issues
afghanistan iraq negotiations wto_regulations wto_talks domestic_laws regional_prosperity debts world_trade_organisation significant	International Politics and Trade
india markets money bb s_rating level five_year_high excuse indian_assets cash	Indian Economy and Finance
exports current_account_deficit europe secretly_happy dublin lower_taxes japanese_yen euro domestic_consumer_demand seven_month_low	Economic Analysis
china developing_nations sub_saharan_countries vogue_landmark textiles textile_workers mauritius textile_quotas agoa united_states	Textile Industry in Developing Countries
oil_ministry iraq output oil_officials foreign_oil_company produce information north_of_the_country s_cabinet offers	Iraq Oil Industry
tough_punishments journalist graft cameroon non_existent_workers retirement prime_minister_ephraim_inoni salary_scandal government_ministries civil_servants	Political Scandal and Corruption
profit shares decline uk_projects west_london claims pre_tax_profits change_in_steel_contractor new_wembley_stadium wembley_stadium	UK Financial News
guilty fraud trial problems worldcom witness books request innocence economist	Legal Proceedings and Economics
oil_firm market surgutneftegaz improvement international_investment_community gazprom half_owned_by_bp renaissance_capital emerging_market_play control	Oil Industry and International Investment
shareholder_transactions settlement earning_growth new_york_attorney_general lawyers trident_funds credit_rating_agencies scandal insurance_broker corporate_cover	Corporate Finance and Legal Issues
compliance_statement air_china internal_controls london_stock_exchange delisting compliance_costs delegation enron_scandal declarations chinese_state_run_banks	Business Compliance and Financial Scandals

Topics	Labels
economic_performance euro_scepticism powerhouse_of_europe largest_port target_of_being_the_most_efficient_economy enlarged_european_union low_skilled_workers long_term_decline global_market efficient_economy	European Economy
spending foreign_exchange_reserves indian_economy service_tax services_sector target personal_tax foreign_investment lower_house_of_parliament resilience	Indian Economy and Finance
world_economic_forum businesses switzerland india new_cases confidentiality ethiopia access_to_treatment non_discrimination social_threats	Global Economic and Health Issues
thailand worst_hit_areas insurers governments southern_asia damage wall_street_journal risks_of_epidemics human_tragedy allianz	Natural Disasters and Insurance

BBC_News-tfidf-lda

Topics	Labels
feta jarvis kronor wembley gaming ericsson yili parking chinese cheese	Unclear
tobacco project three_gorge gorge dam yangtze gold nestle smoking energy	Environmental Projects and Corporate Business
cement cairn ecb boeing druyun myers rossignol federated department_store asbestos	Business and Industry
crude jet barrel oil_price bmw airway opec currency carrier luxury	Economy and Travel
deficit french increased import france trillion benefit gdp looking pension	French Economy
bush qantas nasdaq boeing prime_minister singapore project unemployment_rate ssl technology	Business and Technology News
yukos russian fiat gazprom yugansk russia rosneft court parmalat auction	Russian Business and Legal Issues
wipro wmc foster plastic remittance mexican woman kraft saudi hariri	International Business and Economy
said sale bank firm company share growth year price would	Business and Finance
marsh korea bonus axa construction south_korea insurer shell ministry ntpc	Business and Industry
wall sbc wall_street name illegally prosecutor ebay pension peoplesoft william	Business and Legal Issues
boerse lse deutsche_boerse club ebbers euronext glazer deutsche sullivan manchester	Business and Finance
mci qwest verizon flight cocoa bombardier johnson bafin ukraine saudi	Business and Industry News
game video video_game lira disney alfa renault alfa_romeo romeo call_centre	Entertainment and Automotive Industries
coal pernod laura ifo ashley biogen allied transportation contrack unece	Unclear
turkey metlife wmc chartered standard_chartered fare hiv softbank minute nbc	Business and Finance
rover beer inbev mini bmw fox city chinese umbro greenspan	Business and Economy

BBC_News-tfidf-nmf

Topics	Labels
rate interest_rate price house_price interest house bank housing economy rise	Economics and Housing Market
yukos russian court bankruptcy tax gazprom oil yugansk auction rosneft	Russian Oil Industry
firm share company said foreign investment china new market government	Business and Economics
deutsche boerse deutsche_boerse lse euronext london bid offer stock shareholder	Stock Exchange and Trading

Topics	Labels
sale profit car store year retail said rose euro strong	Economy and Retail
oil crude price barrel opec oil_price supply energy winter crude_oil	Oil Industry
ebbers worldcom sullivan mci former accounting fraud telecom verizon charge	Corporate Fraud and Telecommunications
bank card parmalat creditor credit debt south banking italian korea	Finance and Banking
deficit dollar budget trade euro currency bush export economic record	Economics and Trade
club glazer manchester united manchester_united board bid offer mcmanus man	Football Business
airline air boeing fuel jet flight plane airbus passenger carrier	Air Travel
fiat car alfa auto alfa_romeo romeo italian ferrari engine stake	Italian Automobiles
marsh insurance insurer guilty broker executive sec inquiry insurance_broker spitzer	Insurance Investigation
sri economic disaster people tourism imf reconstruction country lanka aid	Sri Lanka's Economy and Reconstruction
job growth economy unemployment rate economic economist consumer figure creation	Economic Indicators

BBC_News-tfidf-kmeans

Topics	Labels
poverty working_people national_insurance_contributions private_pensions higher_savings uk_state_pension inequality complexity s_pension length_of_residency	UK Pension and Social Inequality
profits revenues chief_executive brands general_motors losses chairman bmw analysts income	Business and Finance
oil_prices economic_growth growth economy consumer_spending unemployment gdp employment inflation government	Economic Indicators
securities_and_exchange_commission internal_controls air_china enron_scandal sec us_stock_market_watchdog foreign_firms sarbanes_oxley_act bank_of_china china_construction_bank	Financial Regulation and Scandals
demand prices brent_crude market production output europe asia barrel libya	Oil Industry
investors equities bubble_burst internet nasdaq public financial_stocks private_placements registration_document dot	Stock Market and Investments
creditors debts banks bankruptcy insolvency chief_executive enrico_bondi apology capital_injection shares	Corporate Financial Crisis
interest_rates consumer_spending dollar euro us_economy federal_reserve unemployment_rate economy labor_department economists	Economics and Finance
sale yukos yuganskneftegas founder mikhail_khodorkovsky political_ambitions auction gazprom russia rosneft	Russian Business and Politics
exports china japan growth world_trade_organisation imports recession domestic_demand recovery records	Economics and Global Trade
housing_market bank_of_england house_prices december slowdown mortgage_approvals rise figures january interest_rates	UK Housing Market and Economy

Topics	Labels
competition company jet_airways oversubscribed shares raise rupees london rivals airline	Airline Industry
euronext deutsche_boerse lse shareholders bid offer chairman talks london_stock_exchange investigation	Stock Market and Trading
euro central_banks dollar currencies bids prosper australian_dollar south_korea election_year uk_businesses	Financial Markets
ukraine elections prime_minister consortium soviet_union president privatisations ryanair krivorizhstal command_economy	Ukrainian Politics and Economy
consolidation takeover shares at shareholders merger sbc_communications bankruptcy deal worldcom	Business Mergers and Acquisitions
fiat distribution loss sales alliance restructuring settlement stake agreement future	Business and Economy
stake malcolm_glazer board offer club shares cubic_expression us_tycoon proposal manchester_united	Business and Sports Finance
news_corp rupert_murdoch chief_operating_officer financial_times investment feature_films activision donkey_kong media_company pac_man	Media and Business
world_economic_forum research china aids business_leaders iraq employers hiv davos political_leaders	Global Politics and Health Issues
fraud collapse charges worldcom executives allegations trial accounting_fraud conspiracy guilty	Corporate Crime
yukos rosneft gazprom russia russian_authorities auction bank_accounts back_tax_bill vladimir_putin tax_evasion	Russian Oil Industry and Tax Evasion
us_firm reports duty_free_stores luxury_goods_group lvmh _chandon_champagne louis_vuitton christian_lacroix non_core_businesses falic_group	Luxury Brands Acquisition
india sri_lanka thailand world_bank indonesia economic_impact rebuilding united_nations economic_growth jobs	Asian Economy and Development
sales demand india profits analyst shares staff results turnover net_income	Business and Finance

Arxiv_Abstracts-tf-lda

Topics	Labels
network learning prototype feature method point image deep kernel local	Machine Learning and Image Processing
method learning knowledge point propose data result entity deep object	Machine Learning and Data Analysis
model learning method image deep data network decision video using	Artificial Intelligence and Machine Learning
model network method object image approach learning task target propose	Machine Learning and Object Recognition
feature learning method model show image semantic adversarial information visual	Machine Learning & Image Processing
learning model approach method data image pose loss problem class	Machine Learning and Image Analysis
learning policy problem reinforcement algorithm method image model reinforcement_learning result	Machine Learning and Image Processing
model image method distribution learning representation propose task proposed data	Machine Learning Techniques

Topics	Labels
image model network method proposed feature data propose generative training	Machine Learning and Image Processing
graph learning neural network node method data representation deep performance	Machine Learning and Data Analysis
network model method learning molecular student result proposed propose data	Machine Learning and Data Analysis

Arxiv_Abstracts-tf-nmf

Topics	Labels
learning reinforcement reinforcement_learning representation machine machine_learning two show problem federated	Machine Learning and Problem Solving
image segmentation using show input based two generate network adversarial	Deep Learning and Image Processing
graph representation node edge pooling graph_pooling representation_learning scene propose graph_neural	Graph Theory and Neural Networks
model existing proposed compression demonstrate propose prediction product performance limited	Machine Learning and Data Compression
object detection object_detection detector saliency query box bounding bounding_box propose	Image and Object Recognition
network neural neural_network convolutional layer architecture result directed novel convolution	Machine Learning and Neural Networks
data distribution synthetic using domain new framework synthetic_data dataset real	Data Generation and Distribution
method show proposed based experiment existing novel result proposed_method propose	Research Process and Methodology
knowledge student teacher distillation different network transfer entity knowledge_distillation propose	Education and Knowledge Transfer
feature information pyramid augfpn attention feature_pyramid experiment different design module	Machine Learning and Image Processing
algorithm policy problem reinforcement function reinforcement_learning result gradient decision agent	Machine Learning and Artificial Intelligence
point cloud point_cloud system segmentation sensor lidar result semantic stereo	3D Sensing and Image Processing
task auxiliary learn auxiliary_task downstream reasoning transfer agent performance curriculum	Machine Learning and Performance Optimization
target pose domain source sar class sar_target labeled learning transfer	Machine Learning and Image Recognition
video temporal attention spatial different information propose novel sequence summary	Video Processing and Analysis
time series time_series forecasting many observed imputation forecast baseline model	Time Series Forecasting
training adversarial generative gan generative_adversarial network loss discriminator performance gans	Machine Learning and AI
approach problem new semantic segmentation existing relation set structured present	Artificial Intelligence and Data Analysis

Topics	Labels
deep deep_learning using learning based clustering deep_neural also performance mechanism	Deep Learning and Performance Optimization

Arxiv_Abstracts-tf-kmeans

Topics	Labels
experiments performance training accuracy state_of_the_art experimental_results reinforcement_learning state_of_the_art_methods deep_learning object_detection	Machine Learning and Performance Evaluation
dataset neurons high_segmentation_performance biomedical_researcher biomedical_images biomedical_image_processing small_datasets myelin_segmentation acquisition_settings annotation_efforts	Biomedical Image Processing
comprehensible_analytical neighborhood tree_structure interpretability local_behavior machine_learning_techniques novel_way support_vector_machine gpx ai_systems	Machine Learning and Data Analysis
c_core_kaggle_competition ship_navigation sar iceberg transfer_learning satellite_imagery data_augmentation subsurface_structures convolutional_neural_network_ geophysics	Machine Learning in Maritime and Geophysics Applications
swappable_components high_quality primates benchmarks state_of_the_art_methods pose artificial_neural_networks texture quantitatively family	Artificial Intelligence and Image Processing
customized_hardware_design convolutional_neural_networks parallel two_stage_detectors maxpoolnms configurable_approach parallelizable_alternative outperforms convolutions accelerating_nms	Computer Vision and Hardware Design
computationally_essential subtle_expressions optical_strain outperform minute_facial_motion_intensities smic_databases spatio_temporal_feature_extraction_approaches two_sets_of_features construct micro_expressions	Facial Expression Analysis and Computation
complexity_of_the_background baseline_system color technology biometric_methods convolutional_encoder_decoder_network variations predicted_image region_of_interest traditional_approaches	Image Processing and Technology
medical_image_processing limitations neural_networks theoretical_foundations image_registration image_segmentation deep_learning rapid_progress popularity applications	Artificial Intelligence in Medical Imaging
memory_friendly data_augmentation inception_score lsun_church higher_resolution low_level_textures training_instability_issues celeba_hq vision_tasks fid	Deep Learning and Image Recognition
s_parameters generator supervised_settings inception_score removal_of_the_instance change_in_the_loss machine_learning_models gan influence_estimation_method gradient	Machine Learning and GAN Models
computer_aided_diagnosis tradaboost_classifiers twitter_spam_detection network_intrusion_detection bootstraped_samples hypothesis target_domain scenarios labeling source_domain	Machine Learning and Cybersecurity
child random_selection_process two_stage_kin_face_generation_model predict generated_images map quantitatively ages deep objective_standards	Artificial Intelligence in Age Progression Techniques

Topics	Labels
visual_information_extraction optimization text_spotting automatic_marking independent_sub_tasks state_of_the_art_methods existing_works intelligent_education public_benchmarks shortage	Artificial Intelligence in Education
_approach active_missions peer_to_peer_negotiations heuristics rl s_deep_space_network s_capacity operational_constraints deep_reinforcement_learning_ untrained_counterpart	Space Missions and Reinforcement Learning
data_variance customers tpg_dnn click_through_rate_ ctr recommendation industry gross_merchandise_volume_ gru shopping_efficiency	E-commerce and Data Analysis
multi_layer_perceptron_ normal_ecg_signals _signals transfer_learning dynamic_features beat_and_rhythm_levels electrocardiography_ accurate_anomaly_detection simple_classifiers order_patterns	Machine Learning in Healthcare
optuna genetic_algorithm sedentary_living_style progress customizable_diet_plans random_search website_framework predict_obesity_levels decision_tree models	Healthcare Technology and Data Analysis
super_resolution training_process proposed_method handwriting_characters generated_images benchmark_data benefits low_resolution_handwriting_character_recognition appropriate attentions	Image Processing and Character Recognition
recognition pixel_level_features offline_mode stroke alignments alignment stroke_constrained_information stroke_level_features decoder hmer	Handwriting Recognition and Analysis

Arxiv_Abstracts-tfidf-lda

Topics	Labels
weakly_supervised weakly acquisition active_learning annotation signature outfit generalization kent tsc	Machine Learning and Data Annotation
color privileged experimental_result partial ship temporal_relation artistic person gaze storage	Unclear
restoration patch caption operation stgcn detail regularization frequency fairness starvqa	Image and Video Processing
cloud point_cloud depth target optical face recognition pair adaptation spatial	3D Imaging and Facial Recognition
concept structured scene_graph auxiliary skeleton entity resolution text greedynms maxpoolnms	Computer Vision and Text Processing
universal internal aerial filter long solved sensor head kernel planar	Engineering Components
translation localization coordinate directed checkpoint feedback license explanation city allowed	Computer Network Routing
imputation curriculum event credit target distillation therapy auxiliary cost character	Education and Therapy Strategies
industrial cell event illumination filter deeper learning_rate regularization fall attentional	Machine Learning and Industrial Processes
label variational representation_learning graph_pooling outlier structured pooling potential cluster energy	Machine Learning and Data Analysis
molecular target body scribble pomo final summary demand schema location	Unclear
representation_learning product adr generator recommender polyp identity transformation text benefit	Machine Learning and Data Analysis

Topics	Labels
image learning method model network data graph approach feature object	Machine Learning and Data Analysis
triplet federated_learning scene_graph prototype relation placement sgg datp capsule object_detection	Machine Learning and Object Detection

Arxiv_Abstracts-tfidf-nmf

Topics	Labels
model data decision tree learning machine deep machine_learning neural time	Artificial Intelligence and Machine Learning
policy reinforcement learning reinforcement_learning algorithm function problem agent bound reward	Reinforcement Learning in Artificial Intelligence
graph node representation neural edge graph_neural learning representation_learning network graph_pooling	Graph Neural Networks
point cloud point_cloud lidar sensor semantic rgb stereo system autonomous	3D Mapping and Autonomous Systems
object detection object_detection saliency detector salient_object salient semantic bounding query	Image Processing and Object Detection
target domain source labeled adaptation transfer data domain_adaptation unlabeled target_domain	Machine Learning and Data Adaptation
image segmentation feature semantic method network information approach input using	Computer Vision and Image Processing
video temporal optical spatial attention face sequence optical_flow frame flow	Video Analysis and Processing
generative gan adversarial training generative_adversarial discriminator gans data generation image	Generative Adversarial Networks (GANs)
scene scene_graph graph reasoning question visual relation structured language task	Visual Data Structuring and Analysis
knowledge student distillation teacher task transfer student_network feature model auxiliary	Machine Learning & Knowledge Transfer

Arxiv_Abstracts-tfidf-kmeans

Topics	Labels
real_world_datasets labels evaluation task bounding_boxes categories subspaces state_of_the_art detections unsupervised	Machine Learning and Data Analysis
researchers weakly_supervised_methods object_boundaries algorithms supervised_methods image_processing semantic_segmentation algorithm natural_language_processing benchmark_datasets	Machine Learning and Image Processing
algorithm deep_network classify renormalization_group initial_condition high_momentum_modes naive_estimates learning_data_set training_times generalization_puzzle	Deep Learning and Classification Algorithms
encoder generative_models gans representation_learning pretrained_models unsupervised_representation_learning unconditional_image_generation self_supervision discriminator image_generation_quality	Deep Learning and Generative Models
cnn improvement experimental_results deep_learning scene color computational_color_constancy attributes convolutional_neural_network_ given_image	Deep Learning and Image Processing

Topics	Labels
transformer state_of_the_art model scene_graphs code attention_mechanism time_complexity video_sequence compositionality deep_learning_models	Artificial Intelligence and Deep Learning
time_series technologies lstm long_short_term_memory_ tabular_data data_privacy security subgroup_differentiation generative_approaches financial_industry	Data Science and Security in Finance
machine_learning deep_reinforcement_learning natural_language_processing adversarial_attacks robustness generalization underlying_spectral_method rigorous_analysis highly_nonlinear_functional word_embedding_method	Artificial Intelligence and Machine Learning
benchmark complexity realistic rogue_gym generalization_ability training_data vital industrial_applications industrial_benchmark variety_of_aspects	Artificial Intelligence Research and Applications
tasks object_detection imagenet image_classification proposed_method learning_to_understand_aerial_images performance student_network teacher_network coco	Computer Vision and Machine Learning
robust key_contributions overview different_tasks survey challenging advances existing_methods computer_vision non_local_neighbors	Computer Vision and Methodologies
inverse_problem reliability training_data redundancy saturation extrapolation experiments dataset parameters simultaneous_detection	Data Analysis and Experimentation
training image_segmentation decision_trees mri convolutional_neural_networks accuracy available code data_augmentation semi_supervised_approach	Machine Learning and Image Processing
sample_complexity convergence state_of_the_art_methods domain_randomization meta_learning q_learning off_the_shelf reinforcement_learning downstream_tasks weights	Machine Learning Techniques
relations entities knowledge_graphs relational_data recommendation state_of_the_art interpretable_model recommender_systems implication_rules algorithms	Knowledge Graphs and Recommender Systems
retinanet fcos inference models object_detection soft_roi_selection object_categories accuracy_trade_off source_code fpn	Machine Learning and Object Detection
uncertainty predictive_accuracy natural_language_processing model performance normalizing_flows computer_vision optical_flow experiments quantifying_uncertainty	Machine Learning and Computer Vision
point_clouds autonomous_driving feature_extraction robotics point_density deep_learning synthetic statistical_properties estimation state_of_the_art	Autonomous Robotics and Deep Learning
scalability adaptation channels integration efficient_rank_one_update model_retraining non_gaussian_bayesian_models gaussian_based graph_based model_change	Machine Learning Models and Techniques
transferability predictions deep_neural_networks method neural_networks evaluate tools overfitting robust_models generation	Deep Learning and Model Evaluation
generator generative_adversarial_networks discriminator realistic_images training_dataset effectiveness gans limited_data generative_tasks generative_performance	Generative Adversarial Networks
theoretical_understanding sample_quality insights gans connection training_stability empirical_performance understanding stability technique	Machine Learning Techniques

Topics	Labels
salient_object_detection saliency_maps fcnn state_of_the_art_approaches semantic_information data_driven feature_redundancy_reduction intrinsic_correlations fully_convolutional_neural_network_ feature_sharing_properties	Computer Vision and Neural Networks
computation approach learning learnable dual_curriculum_scheme learning_problem multicut binary_variables variational_models graph_based_models	Machine Learning and Computation Models
distribution model benchmarks state_of_the_art anomaly_detection molecular_systems method efficiency competitive_results supervision	Machine Learning and Molecular Systems
approaches neural_networks machine_learning classifiers practical_applications random_forest explanations deep_neural_network hyperparameter_optimization decision_tree	Machine Learning Techniques
graph_neural_networks state_of_the_art_performance graph_laplacian gnns graph_pooling graph_data graph_classification graph_pooling_methods effective_message_passing node_embedding	Graph Neural Networks
image usefulness representation architectures convolutional_neural_network segmentation convolution_operation gpc input_image time_consuming	Image Processing and Convolutional Neural Networks
image_generation generative_adversarial_networks super_resolution performance classification proposed_method state_of_the_art_methods experimental_results transformation generated_images	Artificial Intelligence and Image Processing
graph_convolutional_networks gcns flexible competitive feature_propagation_steps large_margin computationally_efficient commercial_smartphones arianna augmented_perception	Artificial Intelligence and Mobile Technology
time_series_classification literature observed_spectrum protocols reflection introduction time_series_benchmark_data_sets ucr public_availability tsc_methods	Data Science and Time Series Analysis
agents data vision effect forecasting target_domain segmentation_results bimodal_resource textual_phrases associations	Computer Vision and Data Analysis
model_compression flops_reduction design_space automl accuracy time_consuming benchmarks support_vector_expansion kernelized_online_learners low_latency_real_time_services	Machine Learning Optimization Techniques
target_policy smoothness mean_square_error off_policy_evaluation offline_data variance reinforcement_learning actor critic value_function	Reinforcement Learning Concepts
unsupervised_learning clustering recall video_segments precision sentiment_analysis product_attributes spray_painting_robots dataset customers	Machine Learning and Data Analysis
reinforcement_learning optimal_policy markov_decision_process optimization algorithms continuous_control objective_function theoretical_contribution novel experience	Machine Learning and Optimization Algorithms
videos github magnitude ground_truth real_world_data objects spatial_network performances cross_link_layers two_stream_flow_guided_convolutional_attention_networks	Computer Vision and Machine Learning
fid gans training generator cgans fr discriminator generative_adversarial_networks_ conditional_generative_adversarial_networks_ contribution	Machine Learning and Generative Models

Topics	Labels
training_set photo_realistic generative_adversarial_network gan_training synthetic_images _image gan generating_out_of_sample_interventional_probabilities automated_sampling causal_bayesian_graph	Generative Adversarial Networks and Image Generation
unlabeled_data target_domain hypothesis transfer_learning source_domain semi_supervised computer_aided_diagnosis extensive_experiments scenarios model_generalization	Machine Learning and Data Analysis
features network experiments high_resolution deep_learning loss_function accuracy remote_sensing spatial_information adversarial_training	Machine Learning and Remote Sensing
accuracy depth effectiveness consistency answer students knowledge_tracing_model predictive_power variation combination_of_models	Educational Data Analysis

ChatGPT_Tweets-tf-lda

Topics	Labels
data chatgpt human think tool answer say tell anything going	Artificial Intelligence Communication
chatgpt model language make language_model thought game new via like	ChatGPT and Language Modeling
chatgpt used application great research way buy based potential trial	AI Technology Use & Potential
chat gpt chat_gpt like ask using get write make use	Chatbot Functions
using read use much artificial artificial_intelligence intelligence take think make	Artificial Intelligence Usage
chatgpt new could asked write like use get people tool	ChatGPT Communication and Usage
chatgpt use good world need may open get free chat	General ChatGPT Operations
chatgpt use using day code know create asked tool time	ChatGPT Usage and Development
get first like one use week new last time chatgpt	General ChatGPT Usage
chatgpt via get right could help know model like text	ChatGPT Communication and Assistance

ChatGPT_Tweets-tf-nmf

Topics	Labels
chat gpt chat_gpt write asked thing asking google bing got	Artificial Intelligence and Search Engines
chatgpt gpt google time via see prompt way say around	Artificial Intelligence Communication
get ask way answer question one google chat bing video	Search and Communication
language model language_model large write text developed work trained generate	Artificial Intelligence and Text Generation
like could feel chute well talk student something essay check	Education and Communication
use book way create tool child help case without write	Educational Resources and Methods
data human people think tool know answer based going talking	Artificial Intelligence and Human Interaction
using image opencv capture notion computer webcam mastering artificial think	Computer Vision and Image Processing
make good would asked criminal need even could work said	General Conversation
new code used latest video see help seo read search	Digital Marketing and SEO

ChatGPT_Tweets-tf-kmeans

Topics	Labels
websites apps launch discord smart_tech implement user_experience bandwagon chatgpt enterprises	Digital Technology and User Experience
chatgpt openai artificial_intelligence google microsoft gpt chatbot technology future bing	Artificial Intelligence and Technology Companies
chat_gpt technical_seo_tips higher google video search_engine_optimization online_marketing seo wordpress_seo_tutorial google_ranking	Digital Marketing and SEO
spit_out entire_internet seconds prompt recreate learning_language_patterns chatbot coherent chatgpt collective_works	Artificial Intelligence and Language Processing
nft mysticism poem openai haiku writing evolve writer attempt logic	Artificial Intelligence and Creativity
language_model chat thread technology example writing possibilities openai world keywords	Chatbot and AI Technology
dogecoin dogecointothemoon apple goodvibesonly chatgpt bbkcf thin_ice crypto doge biden_economy	Cryptocurrency and Social Media Trends
narrow_specific_question experience long_way_to_go wrong chat_gpt generalized fact_checkers factual_data faculty facup	ChatGPT and Fact-Checking
finserv fintech chatgpt enricomolinari marketing govtech parisfinforum ragusosergio wef margaretsiegien	Fintech and Finance Services
ted_hsu ai alamy ai_generated nature scientists photography abstracts aiethics chatgpt	Artificial Intelligence and Photography
shmconverge hospitalistwork chatgpt bottlenecks rhyme ptflow societyhospmed sight life hospitalists	Healthcare and AI Technology
difference chatgpt digital learn pennsylvania innovations spot real tech fake_text	Artificial Intelligence and Technology Learning
underperformance decade tech_news success patients physicians approach applied_innovation innovation_hub innovation_lab	Tech Innovation in Healthcare
future buzz trajectory modern_communication prototype tools public innovation chatgpt ai_fanatics	AI and Technology Innovation
artificialintelligence iclonedna chatgpt bioscience biotech biotechnews biotechnology scientific_research dna cloning	Biotechnology and Genetic Research
options free_trial chatgpt trading buy_signal trades stocks https ideas chart	Stock Market Trading
amount altcoin usdt interest meme bitfinex unknown gems whalealert exchange	Cryptocurrency
artificial_intelligence setting edtech fintech chatgpt education educators flutter iot exploring	Technology and Education
chatgpt language_model tweet concepts artificial_intelligence text large openai topics key_themes	Artificial Intelligence and Language Modeling
reshape perfect_match microsost search_engine risky_moves categories cash_machine productivity_apps meta chatgpt	Technology and AI Development
age age_of_tl leveragetechnology techsavvy tl aitools dw aiforgood didn notionai	Artificial Intelligence and Technology
chatgpt google post insightful hellosurgeai informational_queries aihype rajhans_samdani threaten cherry_picking	AI and Tech Industry
crtvshow intel datatodecisions cxo_talk tech_news awscloud servicenow googlecloudtech chatgpt azure	Technology and Cloud Services

Topics	Labels
tech tensorflow analytics iot openai flutter machine_learning smart_cities python chatgpt	Artificial Intelligence and Technology
chat_gpt language_model write google openai question use people tweet information	Artificial Intelligence and Communication
capture midjourney chatgpt skills learn creating_art ai_arena show daily_tasks compete	Artificial Intelligence and Daily Activities
gpt chat crypto day_sis selling revshark large nft speech sending	Cryptocurrency and Tech Talk
debate tech academic_role interviewed teaching_field digital_sense educational_field alvaropardouy managing_partner ucuoficial	Academic and Technology Careers
chatgpt seo default_behavior answered ditch_search user_behavior changing_fast questions bard queries	ChatGPT and User Interaction
google ad_revenues year adrevenue adsense serp impact direct_answer click people	Online Advertising

ChatGPT_Tweets-tfidf-lda

Topics	Labels
buy trial signal free awaiting based funny access exciting conversation	Marketing and Advertising
time chatgpt finally wild ban know reserve kno wondering come	ChatGPT Interactions
good job teacher chatgpt see chat podcast educational often month	Education and Communication
chat chat_gpt gpt like know ask use want time read	Chatbot Interaction
help used generative enhance district good study verify explanation one	Unclear
written loving maadi good news punishment knowledge event experience research	General Life Events and Experiences
chatgpt model capability growth new share via knowledge access beginning	ChatGPT Development and Promotion
coding show asked use stage gdpr content code timing worry	Programming and Data Privacy
clone story wrong invest copied true question future read transformative	AI Development and Ethics
learn first chatgpt could plant book excel via revolutionize motif	Artificial Intelligence Learning Process
tell created ref power difference get limitation video accurate week	AI Development and Usage
joke wrote teach rivalry article give use used trend weapon	Unclear
implication question definitely cast attacked internet unintentionally crashed follow day	Online Communication Issues
guide chatgpt book summarised bring dna none siri poem supposed	AI and Literature
use via chatgpt heard still found seriously characteristic took ngl	Communication and Perception
current powerful time use write release fiction case industry launchpad	Publishing Industry
ahahaha one buying tried impressed thread really also facing chatgpt	User Experience with ChatGPT
take jordan gpt sure chat meme via easy porky ever	ChatGPT Communication
social_medium social interesting medium damn great think countdown meanwhile well	Social Media Communication
got chatgpt day lot powerful ready need take tech move	AI Technology Progress
built user age twtw idea using super plus year news	Social Media and User Engagement
making love talk bos get launch next era model way	Product Launch and Development
people competition mode impact jobless think change boy tool photoshop	Digital Art and Career Change
using think problem chat_gpt compliment gpt say understanding chat people	Chatbot Communication
look hahahah liability human shit chat lazy able take latest	General Conversation
asked write chatgpt use something data quite make master saying	ChatGPT Interaction and Usage

Topics	Labels
team stuff give service available result made battle mean lawyer	Business and Service
large used language_model language model eye amazing aped eth daily	Artificial Intelligence and Daily Usage
transfer decision drop bing better chat put nerfed nice artificial_intelligence	Chatbot Conversations
help see new girl revolution key profile unvaccinated wrong clearly	Unclear
chatgpt question user asked answer language language_model model asking using	Artificial Intelligence Communication
google chat get chatgpt need gpt chat_gpt anyone big open	Chatbot Technology
long continuation could old spying write game chat great according	Communication and Gaming
via chat black math favorite gpt job claim code bad	ChatGPT-related Vocabulary
even openai paper right weather scientist fantastic powered back fall	Unclear
thanks behaviour racist wayyyyy since training greg thank say chatgpt	Communication and Interactions
new said solution working something use come privacy way chat	Communication and Problem-Solving

ChatGPT_Tweets-tfidf-nmf

Topics	Labels
chat_gpt gpt chat love joke girl shit wrong story got	Chatbot Interactions
chatgpt model language language_model learn knowledge developed trained text make	Artificial Intelligence and Language Modeling
chat created release continuation enters bing come heard chat_gpt complete	Chatbot Development
use example create without heard phone well way else case	General Communication
know need much talking read difference none say let people	Communication
new see read beginning openai blog post model used give	OpenAI Model Updates and Announcements
asked write said something important result saying got tweet quote	Communication/Conversation
like look feel chatbots could chute unvaccinated hype crypto alien	Unclear
ask get google question answer one need great search video	Internet Search and Information Seeking
via capability growth take jordan ever capture opportunity teach back	Personal Development and Learning
good job make said even something cannot understanding without omg	General Conversation
time current powerful data much interesting long prompt future tried	Artificial Intelligence Concepts
free buy trial signal awaiting based trading stock transfer idea	Stock Trading
using human tool think replace people intelligence help power prompt	Artificial Intelligence and Human Interaction

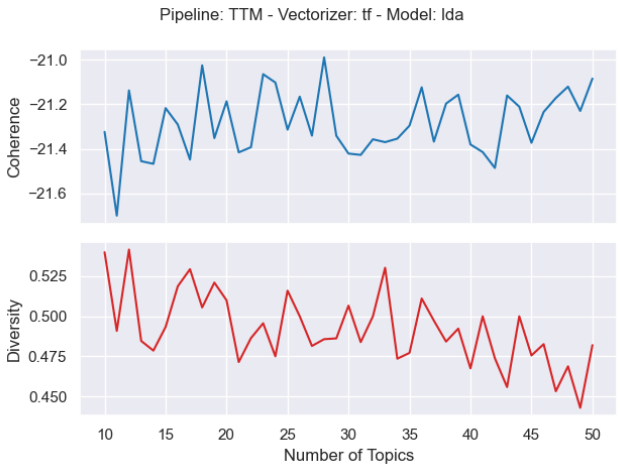
ChatGPT_Tweets-tfidf-kmeans

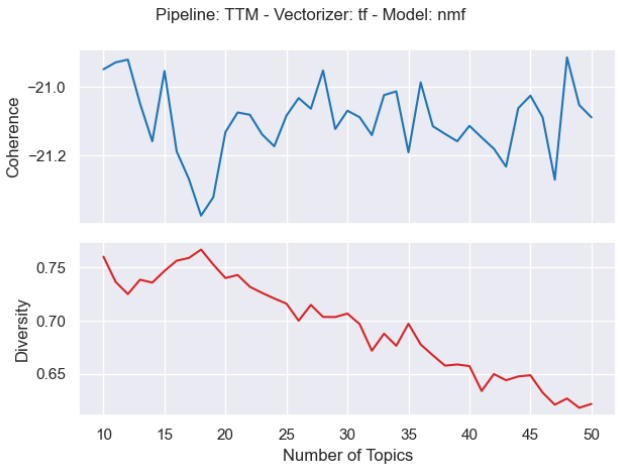
Topics	Labels
artificial_intelligence chatgpt innovation generative_ai technology digital chatbot startups openai nlp	Artificial Intelligence and Technology Innovation
google bard chatgpt microsoft chat_gpt artificial_intelligence seconds compete search_engine openaichat	Artificial Intelligence Companies
chat_gpt write wrong thegoldsuite take_over people information jokes jordan_peterson generalized	ChatGPT and Social Interaction
job midjourney hilarious chatgpt eyes professional aiart discussion unisex_softstyle_tshirt etsy	ChatGPT and Professional Discussions
openai chatgpt chatbot world chat_gpt microsoft work plants officially_released name	Artificial Intelligence and Technology

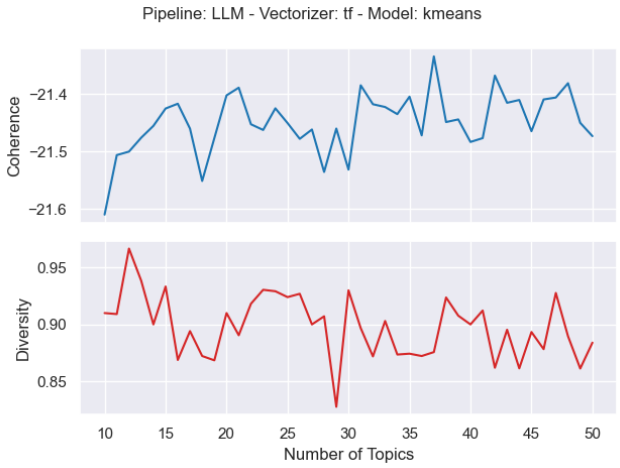
Topics	Labels
use chatgpt chat_gpt trending excited account michael_porter david_espinoza tailings_dams professionals	ChatGPT Discussions and Users
keywords dalle thread gpt join midjourney imagine text writing example	Artificial Intelligence and Writing
thoughts scientists chatgpt julainespeight long_discussion professional_help thinker friends game_changer article	ChatGPT Discussions and Contributions
prompt chatgpt model chat_gpt generate output input current_resources developers tokens	Artificial Intelligence Development
question answer chat_gpt gpt_chat make_money minds important start_building lewishowes stanleymasinde	Artificial Intelligence and Entrepreneurship
talk girl ser become_irrelevant evolve chat_gpt online_business paid_version ways_to_earn_money groove_digital	Chatbot and Online Business Development
language_model chatgpt tweet concepts chat_gpt large key_themes main_topics chatbot topics	ChatGPT and Topic Modeling
amount dump value_in_usd transfer unknown whale_alert whalealert meme pump crypto_contract	Cryptocurrency Transactions
bing video chatgpt move microsoft search_engine google sundarpichai seo content_creation	Digital Marketing and Search Engine Optimization
airdrop future nft crypto chatgpt btc eth magic nostr bonk	Cryptocurrency and ChatGPT
compliment koushikjoshi chat_gpt zoom_pro facup fact_checkers factual_data faculty failed fact	Communication and Verification
chatgpt write gpt people poem power prompts new teachers gpt_ai	Artificial Intelligence and Creativity
tech education chatgpt artificial_intelligence machine_learning iot edtech fintech analytics python	Technology and Education
chat gpt catalinmpit influence base god_bless story crypto ask excel	Social Media Interaction

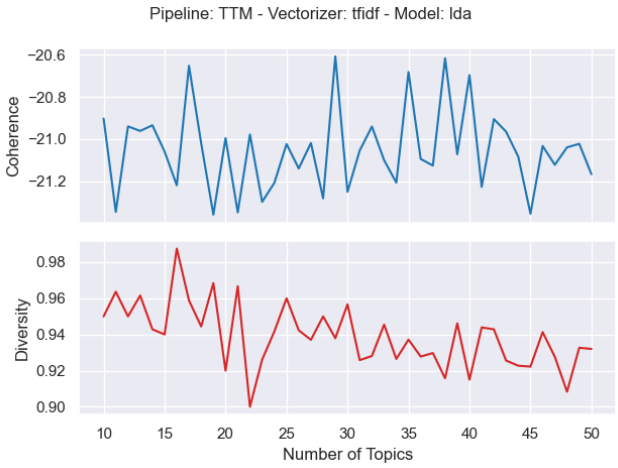
Appendix B: Quantitative Evaluation Plots

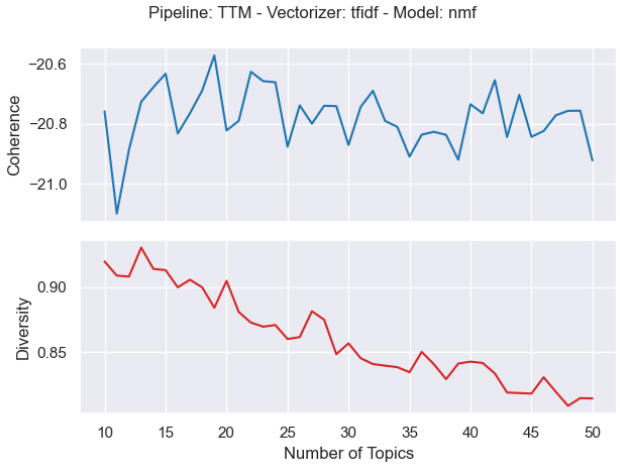
BBC Business News

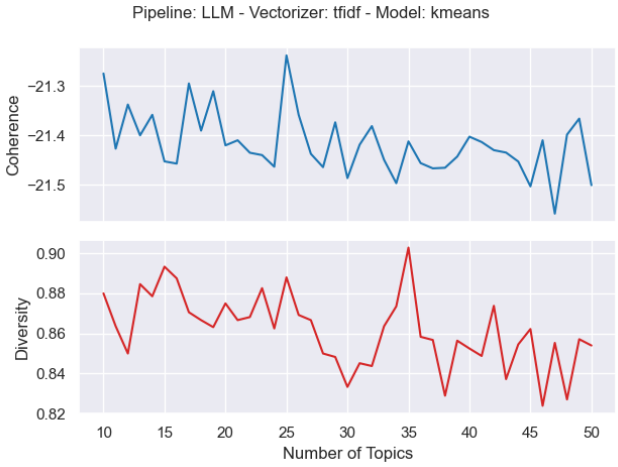




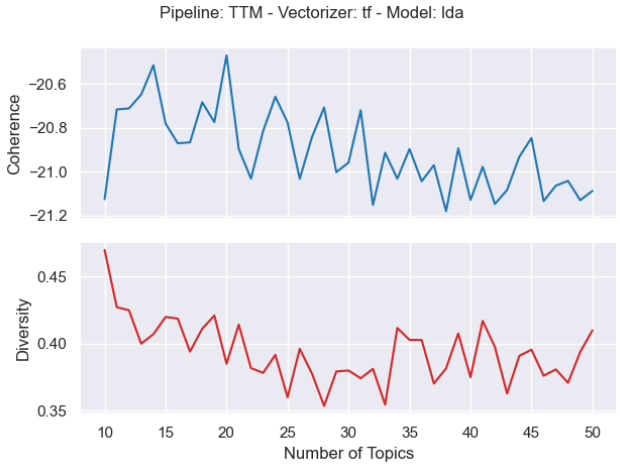


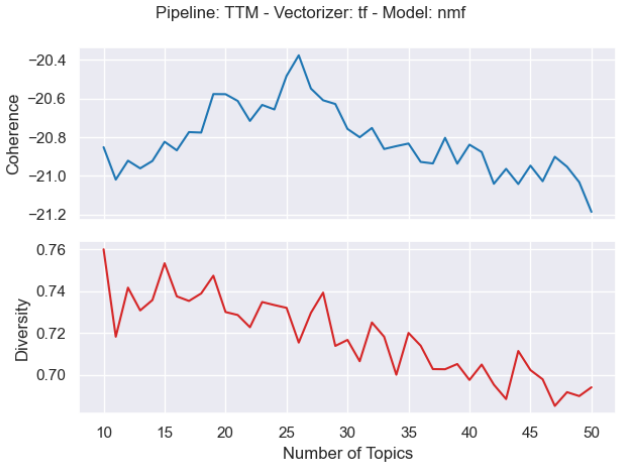


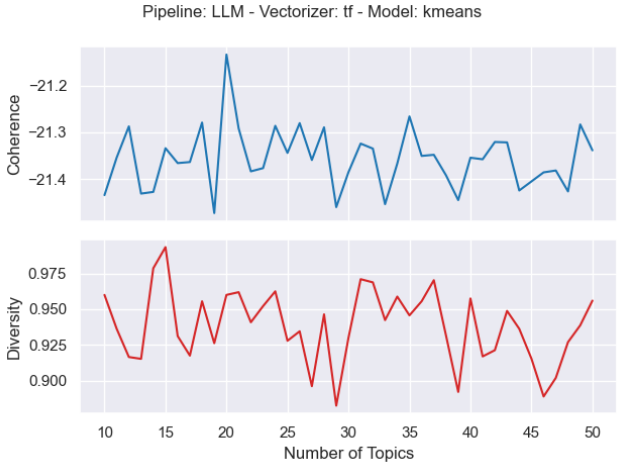


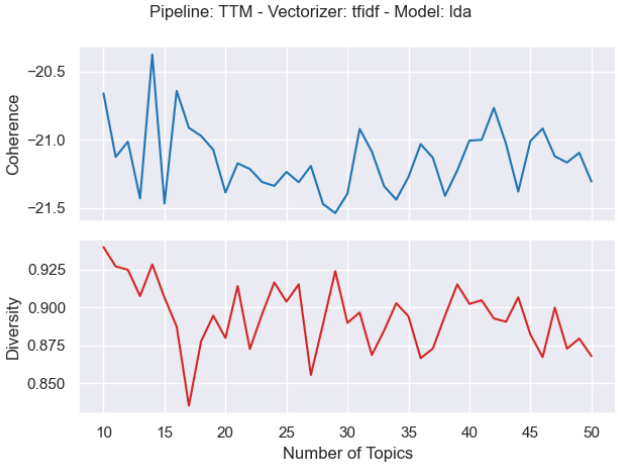


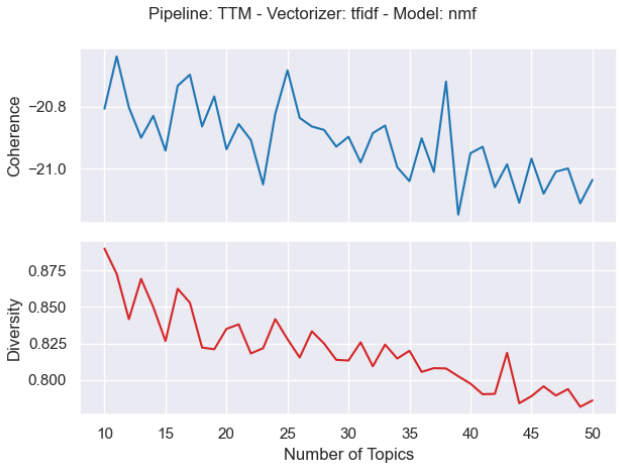
Arxiv Abstracts

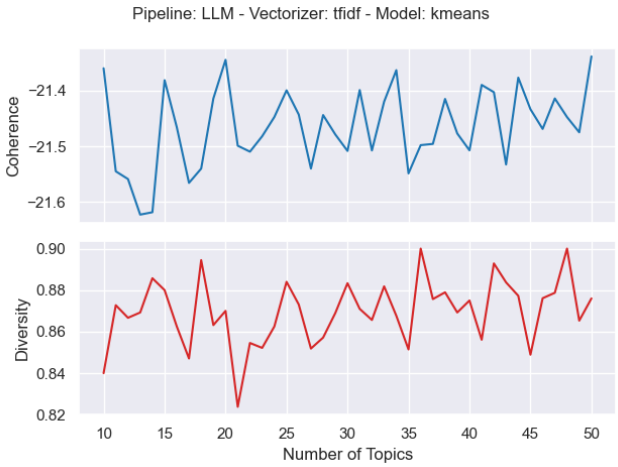




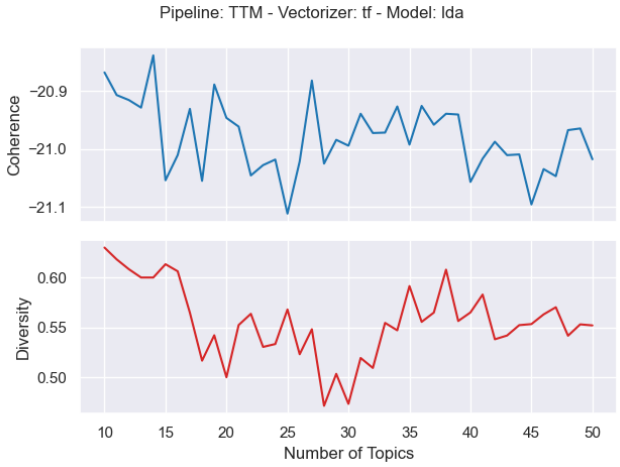


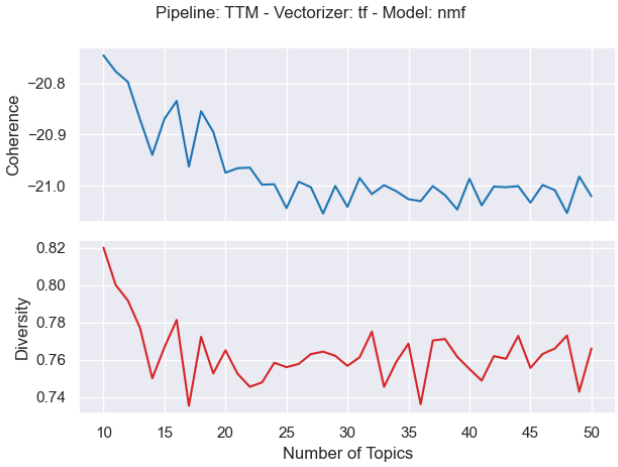


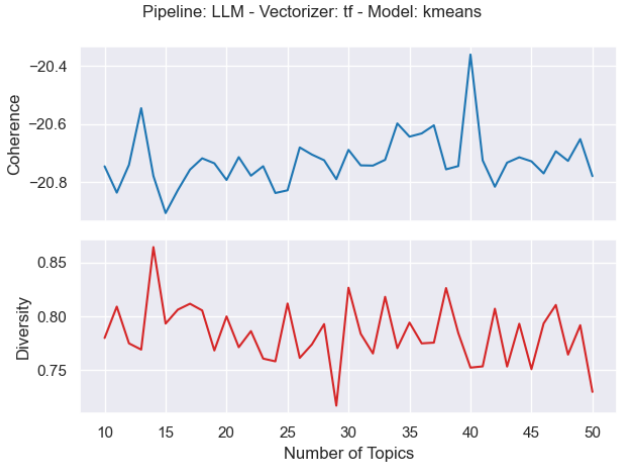


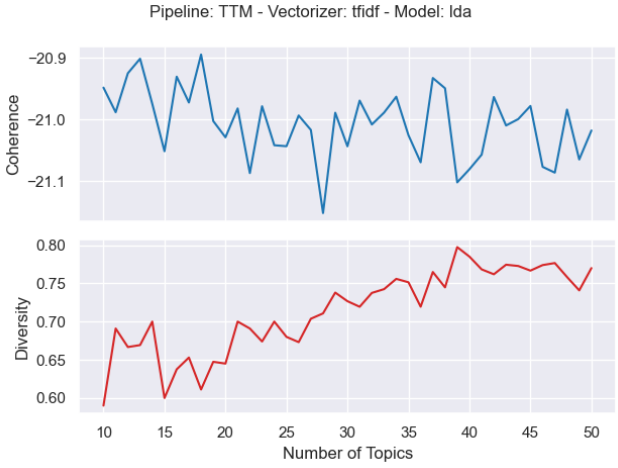


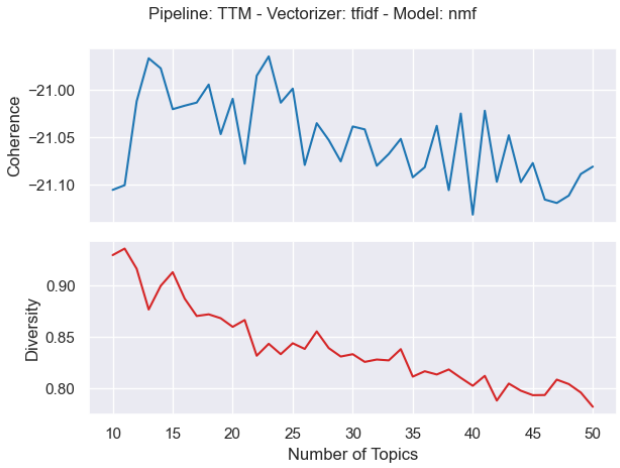
ChatGPT Tweets

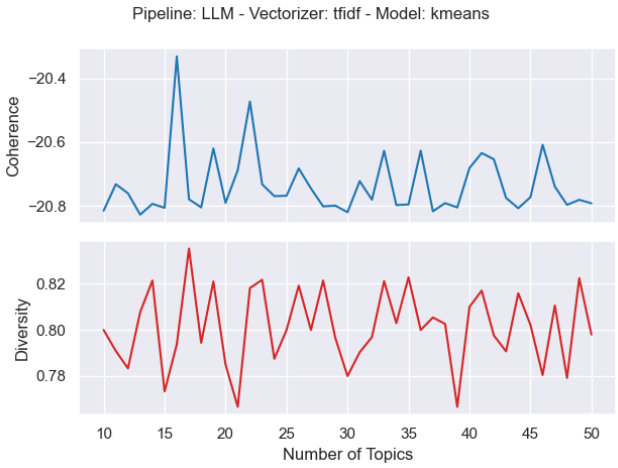












Appendix C: Prompts

Keyword Extraction

Please extract the most relevant keywords that represent the main topics or concepts discussed in the text below. The text is a {context}. I want to identify the key themes and concepts discussed in the text to gain a quick understanding of its content. Provide a list of keywords in a comma-separated format. Prioritize nouns and key phrases over common words or stop words. The extracted keywords should be single words or short phrases that accurately capture the essence of the text.

Example keyword list format:
keyword1, keyword2, keyword3, keyword4, keyword5, keyword6

Text: ###
{document}
###

Word Intrusion

Please assess this list of ten words generated by a topic model. The words come from a collection of {context}. For each list, your objective is to determine the words that do not fit with the others based on their relationships.

Example 1:
Word List: baby crib diaper beer pacifier cry fridge
In this example, the least related words are 'beer' and 'fridge' because it is unrelated to infants, unlike the other words which are closely associated with infants.

Example 2:
Word List: hard_drive motherboard video_card processor ram usb_key
In this case, the least fitting word is 'usb_key' because it stands out as the only item that is not an internal component of a computer.

Respond only with the number of words that do not fit.
Words:
###
{words}
###

Coherence Rating

Please assess the degree of relatedness among groups of words on a 3-point scale. The words were generated by a topic model and come from a collection of {context}. There are the following rating options: Very Related, Somewhat Related, Not Very Related

Choose 'Very Related' when most of the words within a group exhibit a clear and easily describable relationship. You should be able to readily articulate how these words are connected.

Very Related Example 1:
Words: dog cat hamster rabbit snake
Relationship: 'Pets' (An obvious way to describe the relationship)

Very Related Example 2:
Words: brushwork canvases expressionism cubism modernism curators abstract_expressionism national_gallery_of_art museum fossils
Relationship: 'Art' (Although not all words are directly related to 'Art', the overall connection is clear)

Choose 'Somewhat Related' when the words within a group are loosely connected, but there might be a few ambiguous, generic, or unrelated words present.

Somewhat Related Example 1:
Words: computer video new plug screen model
Some words are generic, and the relationship between them is not as strong. Some words may appear to be more closely related than others.

Somewhat Related Example 2:
Words: dog ball pet receipt pen
While some words may seem related, not all of them share a strong connection.

Choose 'Not Very Related' when the words in a group lack any obvious relationship, and it would be challenging to describe how they are connected.

Not Very Related Example:
Words: dog apple pencil earth computer

Answer only with the rating.

Words:
###
{words}
###

Topic Labels

Please label this list of ten words generated by a topic model with a descriptive topic. The words come from a collection of {context}. For each list, your objective is to describe the words with a label that captures the relationship of the words. If there is no clear relationship between the words, label the words with 'Unclear'. Answer only with the topic.

Example 1:
Words: dog cat hamster rabbit snake
Topic: Pets

Example 2:
Words: brushwork canvases expressionism cubism modernism curators abstract_expressionism national_gallery_of_art museum fossils
Topic: Art

Example 3:
Words: dog apple pencil earth computer
Topic: Unclear

Output Format:
Topic

Words:
###
{words}
###

Appendix D: Python Code

The Python code can be found in the following GitHub repository:

https://github.com/marco507/Trade-Offs-Between-Large-Language-Models-and-Traditional-Statistical-Algorithms-for-Topic-Modeling