Evaluating Document Clusters through Human Interpretation

Anton Eklund

Doctoral Thesis, March 2025
Department of Computing Science
Umeå University
Sweden

Department of Computing Science
Umeå University
SE-901 87 Umeå, Sweden

anton.eklund@cs.umu.se

ISBN 978-91-8070-646-9 (print)
ISBN 978-91-8070-647-6 (digital)
ISSN 0348-0542
UMINF 25.03

Cover: "*Human interpreting clusters.*" News clusters visualized using Nauck's nebula, and Anton's color-quantized eye using K-Means clustering.

Printed by Scandinavian Print Group, Umeå University, 2025

Abstract

Document clustering is a technique for organizing and discovering patterns in large collections of text, often used in applications such as news aggregation and contextual advertising. An example is the automatic grouping of news articles by theme, which is the focus of this thesis. For a clustering to be successful, typically the resulting clusters need to appear interpretable and coherent to a human. However, there is a lack of efficient methods to reliably assess the quality of a clustering in terms of human-perceived coherence, which is essential for ensuring its usefulness in real-world applications.

To address the lack of evaluation methods for document clustering focusing on human interpretation, we introduced Cluster Interpretation and Precision from Human Exploration (CIPHE). CIPHE tasks human evaluators to explore document samples from a cluster and collects their interpretation. The interpretation is collected through a standardized survey and then processed with the framework metrics to yield the cluster precision and characteristics. This thesis presents and discusses the development process of CIPHE. The feasibility of performing the exploratory tasks of CIPHE in a crowdsourcing environment was investigated, which resulted in insights on how to formulate instructions. Additionally, CIPHE was confirmed to identify characteristics other than the main theme such as the negative emotional response.

CIPHE was paired with a standard clustering pipeline to evaluate its capabilities and limitations. The pipeline is widely applied for its adaptability and conceptual simplicity, and also being part of the popular topic model BERTopic. The empirical results of applying CIPHE suggest that the pipeline, when integrated with a Transformer-based language model, generally yields coherent clusters.

Additionally, topic models have a similar aim as document clustering which is to automate the corpus processing and present the underlying themes to a human. Topic modeling has rich research on the human interpretation of topic coherence. In the thesis, the human interpretation collected with CIPHE was related to established research in topic coherence. Specifically, the human interpretation collected with CIPHE was used to highlight limitations with the keyword representations that topic coherence evaluation relies on.

Populärvetenskaplig sammanfattning

Klustring av dokument är en teknik för att organisera och upptäcka mönster i stora textsamlingar (korpus), ofta använd i tillämpningar såsom nyhetsaggregering och kontextuell annonsering. Ett exempel är automatisk gruppering av nyhetsartiklar efter ämne, vilket är fokus för denna avhandling. För att en klustring ska vara användbar behöver de resulterande klustren vanligtvis vara möjliga att tolka och uppfattas koherenta av en människa. Det råder dock brist på effektiva metoder för att på ett tillförlitligt sätt bedöma kvaliteten på en klustring i termer av mänskligt upplevd koherens, vilket är avgörande för att säkerställa dess användbarhet i verkliga tillämpningar.

För att addressera bristen på utvärderingsmetoder för dokumentklustring med fokus på mänsklig tolkning har vi introducerat ramverket Cluster Interpretation and Precision from Human Exploration (CIPHE). CIPHE ger mänskliga utvärderare uppdraget att utforska stickprov av dokument från ett kluster och ge sin tolkning. Tolkningen samlas in genom en standardiserad undersökning och bearbetas sedan med ramverkets metriker för att estimera klustrets precision och karaktärsdrag. Denna avhandling presenterar och diskuterar utvecklingsprocessen för CIPHE.

CIPHEs lämplighet att användas i en crowdsourcing-miljö undersöktes, vilket resulterade i insikter om hur instruktionerna givna till deltagarna bör formuleras. Dessutom bekräftades att CIPHE kunde identifiera klustrens karaktärsdrag utöver ämne, såsom negativa emotionella reaktioner.

CHIPE användes under arbetets gång tillsammans med en populär klustringspipeline för att utvärdera ramverkets möjligheter och begränsningar. Pipelinen är populär på grund av sin anpassningsbarhet och konceptuella enkelhet samt att den är en del av den populära topic modellen BERTopic. Slutsatserna från de empiriska resultaten vid tillämpning av CIPHE indikerar att pipelinen, när den kombineras med en Transformerbaserad språkmodell, i de flesta fall producerar sammanhängande kluster. För att förbättra kvalitén på kluster i specialiserade uppgifter som t.ex. kontextuell annonsering, så bör vektoriseringsprocessen anpassas mot att skapa kompakta och separerade kluster i vektorrummet.

Ett närliggande forskningsfält kallas för Topic Modeling. Topic-modeller har ett liknande syfte som dokumentklustring i att automatisera bearbetningen av ett korpus och sedan presentera de underliggande ämnena för en människa. Topic-modellering har djupgående forskning kring mänsklig tolkning av topic-koherens. I denna avhandling relateras arbetet med mänsklig tolkning genom CIPHE till etablerad forskning inom topic-koherens. Specifikt används den mänskliga tolkning som samlats in med CIPHE för att belysa begränsningar med de nyckelordsrepresentationer som utvärdering av topic-koherens bygger på. En konklusion från avhandlingen är att de begränsningar som belystes indikerar att utvärderingen av topic-koherens bör skifta till att fokusera mer på människocentrerade metoder, såsom CIPHE.

Detta arbete betonar vikten av mänsklig utvärdering av system som använder avancerad AI. Kostnaden för att utvärdera modeller ökar i takt med att systemen blir alltmer kapabla. Därför är det avgörande att utveckla effektiva metoder för mänsklig utvärdering så att AI kan användas för mänskligheten bästa.

Preface

This doctoral thesis contains the following papers.

Paper I Anton Eklund and Mona Forsman. Topic Modeling by Clustering Language Model Embeddings: Human Validation on an Industry Dataset, *Empirical Methods in Natural Language Processing 2022: Industry*, Abu Dhabi, 2022.

Paper II Anton Eklund, Mona Forsman and Frank Drewes. CIPHE: A Framework for Document Cluster Interpretation and Precision from Human Exploration, *The 4th workshop on Natural Language Processing for Digital Humanities*, Miami, 2024.

Paper III Anton Eklund, Mona Forsman and Frank Drewes. Comparing Human-Perceived Cluster Characteristics through the Lens of CIPHE: Measuring Coherence beyond Keywords, *Journal of Data Mining and Digital Humanities, NLP4DH*, 2025.

Paper IV Anton Eklund, Mona Forsman and Frank Drewes. An Empirical Configuration Study of a Common Document Clustering Pipeline, *Northern European Journal of Language Technology, Volume 9*, 2023.

Paper V Anton Eklund, Albin Nordström, Frank Drewes and Mona Forsman. Industry Quality Control for Efficient Continuous Human Validation of Deployed Text Classification Systems, *Submitted to ACL 2025: Industry*.

In addition to the papers included in this thesis, the following publications were published within the studies but not contained in this doctoral thesis.

Paper VI Arezoo Hatefi, Anton Eklund and Mona Forsman. PromptStream: Self-Supervised News Story Discovery Using Topic-Aware Article Representations, *LREC-COLING 2024*, Torino, 2024.

Paper VII Hannah Devinney, Anton Eklund, Igor Ryazanov and Jingwen Cai. Developing a Multilingual Corpus of Wikipedia Biographies, *Recent Advances in Natural Language Processing 2023*, Varna, 2023.

The doctoral student was employed at Codemill and Aeterna Labs (previously Adlede) for the duration of their studies. Funding for the doctoral studies was provided by the Swedish Foundation for Strategic Research (ID19-0055). Additional resources for experiments were obtained from the High Performance Computing Center North (HPC2N) and WARA Media and Language.

Acknowledgments

"*You shall know a word by the company it keeps*" – J.R. Firth (1957)

The meaning of a word can be understood by its surroundings. For me, the people who surround me give me meaning, and I can only hope that I am an average of my social context in terms of intelligence, curiosity, kindness, and empathy. Common to everyone who has been with me on this journey is their unconditional support and friendship.

First, I want to thank my splendid supervisors, Frank Drewes and Mona Forsman, for their mentoring and support. I hope our discussions on the endless topics of human interpretation and clustering were as rewarding for you as they were for me. Thank you, Mona, for always being available to conceptualize ideas on the whiteboard and for matching my energy in discussions on matters of science and life. Thank you, Frank, for giving me the space to seed and grow my ideas, and for being a dependable source of priceless advice.

Thank you to my colleagues at Aeterna — Mona, Benjamin Björklund, Kabir, Jon, Benjamin Nauck, Chris, Adrian, Dusan, Sofia, Rickard, and Johanna — for providing me with a problem to study, an environment to grow, but most of all, a place to belong. I could never have made it through without your support and camaraderie.

Thank you to all the people in the Foundations of Language Processing group at Umeå University. You are the most empathetic and selfless group of people I have met. To the leadership, — Frank, Johanna, Henrik, Martin, and Rahil — and to the current and former PhD students, — Jingwen, Emil, Anna, Igor, Lena, Arezoo, Hannah, Adam, and Willeke — it was a joy to share this time with you, and I hope we stay in touch in the future.

Thank you to all my friends and relatives who helped me find perspective beyond conducting experiments, writing papers, and stressing over deadlines.

Thank you to my family — Anders, Maria, and Sanna Andrea — for always believing in me and for providing a safe base for me to grow. It means everything to know that I have a home to return to, no matter what journey I choose to embark on.

Thank you Hayoung for being my greatest cheerleader, for always being there to listen to my worries, and for caring about my well-being. I am happy to have shared this chapter of my life with you — with more to come!

Glossary

Aeterna Labs	Industry partner in Contextual Advertising
AMI	Adjusted Mutual Information Index (extrinsic metric)
ARI	Adjusted Rand Index (extrinsic clustering metric)
BERT	Bidirectional encoder representations from transformers (language model)
BERTopic	Topic model based on the clustering pipeline
CIPHE	Cluster Interpretation and Precision from Human Exploration (evaluation framework)
cluster	A group of similar documents
clustering model	A model that clusters similar documents in a corpus
corpus (plur. corpora)	A collection of text documents
C_V	Topic coherence metric
C_{NPMI}	Topic coherence metric
C_{UMASS}	Topic coherence metric
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise (clustering algorithm)
Intrusion	Human interpretation collection method
investigator	The entity performing the clustering and analysis
K-Means	Centroid-based clustering algorithm
KWM	Keyword-based methods (human interpretation collection methods based on keywords)
LLM	Large Language Model
NLP	Natural Language Processing
participant	A crowdsource participant evaluating cluster samples
PCA	Principal Component Analysis (dimension reduction)
STELLAR	Systematic Topic Evaluation Leveraging Lists of ARticles (corpus browsing tool)
T5	Text-to-Text Transfer Transformer (language model)
theme	A subject or topic of discussion as understood by humans
topic	A set of k descriptive keywords
topic coherence	Collection of metrics from topic modeling literature
topic model	A model that produces topics that describes a corpus
UMAP	Uniform Manifold Approximation and Projection (dimension reduction algorithm)
unsupervised	Without the use of labeled data

Contents

1 Introduction **1**
- 1.1 Contextual Advertising . . . 3
- 1.2 Research Aims . . . 3

2 Background **5**
- 2.1 Document Clustering . . . 5
  - 2.1.1 Vectorization . . . 7
  - 2.1.2 Dimension Reduction . . . 7
  - 2.1.3 Clustering . . . 8
  - 2.1.4 Evaluating Clusters . . . 9
- 2.2 Topic Modeling . . . 10
  - 2.2.1 Topic Model Evaluation . . . 10

3 Methodological Overview **13**
- 3.1 The Clustering Pipeline . . . 13
  - 3.1.1 BERTopic . . . 14
- 3.2 Datasets . . . 14
- 3.3 Cluster Sample Validation Process . . . 16
  - 3.3.1 STELLAR . . . 17
  - 3.3.2 CIPHE . . . 17

4 Summary of Papers and Contributions **21**
- 4.1 Paper I . . . 22
- 4.2 Paper II . . . 22
- 4.3 Paper III . . . 24
- 4.4 Paper IV . . . 26
- 4.5 Paper V . . . 27

5 Synthesis and Discussion **29**
- 5.1 Evaluating the Cluster Sample Validation Process . . . 29
- 5.2 Topic Coherence Examined through CIPHE . . . 31
- 5.3 Evaluating the Clustering Pipeline . . . 32
- 5.4 Future Work . . . 34

6 Conclusions **35**

Chapter 1

Introduction

Document clustering is a technique for organizing large text corpora into groups of similar documents. It has diverse applications but is particularly useful in scenarios where the goal is to uncover underlying structures within a dataset or adapt to data in an unsupervised setting. A common example is the automatic grouping of news articles by theme. The volume of daily news articles spanning different themes is too large to efficiently comprehend without algorithmic assistance. A reliable clustering algorithm saves considerable effort in the discovery of new trends and themes in the data.

The complex and nuanced nature of human communication makes annotating texts with labels and categories for general application difficult. What determines if a clustering is successful largely depends on the algorithm application purpose, and whether humans are able to interpret coherent patterns in the clusters. Consider the example in Table 1.1 with five news articles that are clustered together. All five articles could collectively be viewed as a coherent *Weather* cluster. However, depending on the requirements of the cluster model application, this could be viewed as two themes, *Hurricanes* and *Forecasts*. If two themes are wanted, the clustering model may need to further divide the cluster. To reliably evaluate clusterings of text documents, there needs to be structured methods for data collection which considers human interpretation and are adaptable to specialized application settings.

Topic modeling is a field closely related to document clustering but has a slightly different aim. Its primary goal is to uncover and explain the main themes within a corpus to a human, usually representing them as coherent sets of keywords i.e. a *topic*. A document clustering model can be turned into a topic model by algorithmically annotating clusters with keywords. The topic modeling field has rich research regarding the human interpretation of *topic coherence*, meaning the perceived coherence of the keyword sets representing themes. However, the topic coherence metrics have been criticized for being overly reliant on word co-occurrence which does not capture the complex nuances of human perception (Doogan and Buntine, 2021; Hoyle et al., 2021).

Topic 0	spells, sunny, hurricane, forecast, cloudy, highs, rain, temperatures, weather
id: 777	1st of 2022, Hurricane Agatha heads for Mexico tourist towns
id: 2510	Hurricane Agatha is first named storm of Atlantic season after hitting Mexico
id: 2197	Met Office gives Scotland weather update for Queen's Platinum Jubilee week
id: 4899	The wait is over, Met Office has revealed Platinum Jubilee weather forecast
id: 5770	London weather: Exactly when temperatures in capital are expected to soar

Table 1.1: A *Weather* cluster that can be divided into *Forecasts* and *Hurricanes* depending on the model purpose. From Paper I (Eklund and Forsman, 2022).

This thesis focuses on the intersection between document clustering and topic modeling, and aims to examine the topic coherence metrics through the interpretation of the underlying document clusters.

Document clustering models do not naturally extract keywords so the topic modeling coherence evaluation can not be applied directly. Thus, the work in this thesis is motivated by a need for collecting human interpretation of clusters without the use of keywords. This thesis centers around what was defined as a *cluster sample validation process* (Section 3.3) that puts emphasis on trusting human judgment to evaluate groups of texts. The process can be summarized as following. Given a sample of documents from one cluster, a human evaluator is assumed to be able to: a) contextualize the documents in the sample and decide what they have in common, b) remove any articles not fitting to this context, and c) name the context.

This thesis reports on the development for turning this process into the resulting framework *Cluster Interpretation and Precision from Human Exploration* (CIPHE). CIPHE converts the qualitative process of manually inspecting samples from clusters into a structured framework for expert evaluation or crowdsourcing.

The development of CIPHE highlights many interesting discussions on the notion of human-interpreted coherence and contradictions around using gold labeling for evaluation of document clusters. Through collecting document cluster data with CIPHE, we critically analyze the current topic coherence paradigm where we agree with criticism of the keyword co-occurrence meth-

Terminology The topic modeling field often uses the word **topic** for describing the keyword representation, and in extension **topic coherence** refers to the coherence of these keywords. In this thesis, a subject or topic of discussion as it is understood by humans is referred to as a **theme**. E.g. the interpreted cluster theme is *Football* and the topic is ⟨`football`, `player`, `premier`,..., `league` ⟩.

ods used in topic modeling evaluation. Finally, the clustering pipeline used throughout the thesis is discussed to analyze the value of the information that can be gained from using CIPHE and summarize what has been learned about clustering with Transformer-based language model embeddings.

1.1 Contextual Advertising

The application environment of a clustering system defines its limitations and requirements. This thesis is set in the environment of contextual advertising in close collaboration with the industry partner Aeterna Labs.[1] Contextual advertising is a form of digital advertising where the targeting is derived from the website page content, as opposed to personal targeting that uses personal information through website cookies. For example, an article about sunny weather gets a matching advertising of sunscreen instead of an office chair that the website visitor recently clicked on.

Contextual advertising in news media places certain requirements on the content analysis. The work in this thesis mainly focuses on news websites and larger blogs, excluding social media, forums and commercial websites. Emphasis is put on the categorizations having high precision since advertisers need to match their ads with the expected content of the category. As an example, the category *Sunny weather* should not contain articles on the legal dispute between Prince Harry and the owner of *The Sun*. Thus, the importance lies in correctly categorizing semantically similar articles, i.e. the articles should discuss similar themes. Additionally, news stories change rapidly so the categories should stay updated with the latest developments in the world. This is where clustering allows for the discovery of new stories and can adapt to changes in the older ones. However, there is a lack of methods for ensuring that a clustering model performs adequately when included in a production system. Thus, the research in this thesis addresses a need for clustering evaluation methods that can be adapted to specialized applications such as contextual advertising.

1.2 Research Aims

Applications such as contextual advertising relies on the resulting clusters from automatic algorithms to be interpretable by humans. The complex nature of human communication demands examining a cluster beyond a general theme, which takes into account human attributes such as emotion or opinion. This results in the following research aims:

AIM1 Design and evaluate a versatile framework around the cluster sample validation process.

[1] https://aeternalabs.ai/

AIM2 Compare the output from such a framework with keyword-based methods from topic coherence research.

AIM3 Investigate the possibilities and limitations of a clustering system employing a general pipeline including Transformer-based language models in the setting of contextual advertising.

Each of these aims are discussed in their own section in the chapter on Synthesis and Discussion (Chapter 5).

Chapter 2

Background

This thesis mainly builds on two research fields. The first is document clustering that provides the theory for grouping similar text documents (Section 2.1). The second is topic modeling evaluation which provides a rich literature regarding linking human interpretation with the algorithmic output of topic models (Section 2.2).

2.1 Document Clustering

Document clustering is the process of grouping a collection of documents into clusters based on their content similarity, such that documents in the same cluster are more similar to each other than to those in other clusters (Aggarwal and Zhai, 2012). Document clustering can be used in a wide range of fields including information retrieval, recommendation systems, and news trend discovery. The main case related to this thesis is using clustering to discover themes and events in large news corpora.

A general clustering pipeline contains the steps pre-processing, vectorization, dimension reduction, clustering, and analysis as seen in Figure 2.1. The pre-processing step usually involves data pruning by removing documents that should not be part of the analysis. Then, for a computer to process the natural language in text, it needs to be transformed to a numerical representation in the vectorization step (Section 2.1.1). Next, the vectors are processed to reduce noise and reveal latent features in an optional step of dimension reduction (Section 2.1.2). Lastly, a clustering algorithm is applied to the resulting vectors which assign the same label to documents that have close distances to each other (Section 2.1.3). The analysis step, where the model is evaluated or the results are used for informed decision-making, is where this thesis make its contribution. Following are the background for each of the components related to the thesis and news clustering.

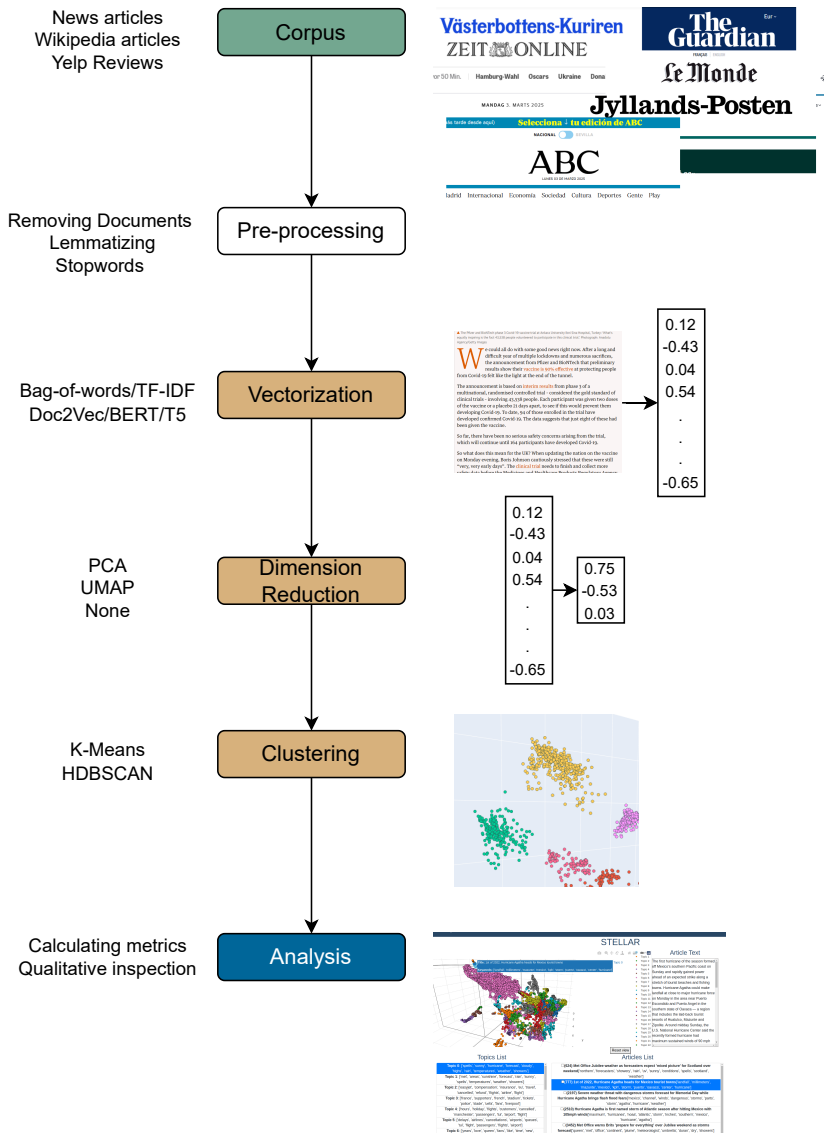


Figure 2.1: An overview of the general clustering pipeline used in this thesis. The steps marked in yellow are the mainly discussed steps. The analysis part marked in blue is where this thesis makes its contributions.

2.1.1 Vectorization

Documents are prepared for algorithmic analysis by mapping them to vectors. Traditionally, vectorization was done through the vector space model (Salton et al., 1975), which uses count-based methods such as Bag-of-Words (Manning et al., 2008) and term frequency-inverse document frequency (TF-IDF, Salton and McGill (1983)). Bag-of-Words is a basic form of vectorizing by counting the number of occurrences of each word in a document. Each dimension in a vector corresponds to a word in the corpus vocabulary, with its value representing how many times that word appears in a document. Since most words do not appear in most documents, this results in a sparse vector. A statistical evolution of this method is TF-IDF where the occurrence of common words, such as "and" or "it", have reduced weight to give room for more indicative words for the specific document.

The main issue with these methods is that word-order and grammar are disregarded. Another issue prominent within clustering is that the sparse vectors give rise to high-dimensional feature spaces, leading to increased computational complexity and reduced efficiency in distance-based similarity measures (Zimek et al., 2012). To address such issues, neural methods have largely replaced traditional count-based methods by mapping words to more condensed vectors and taking the word context into account. The Word2Vec algorithm (Mikolov et al., 2013) is based on a neural network training to predict the word based on the surrounding words in a context window[1]. One effect is that words that are used in similar contexts get a more similar vector representation, an effect highly beneficial for many Natural Language Processing (NLP) tasks including clustering. The document-scope extension of this technique is Doc2Vec (Le and Mikolov, 2014) which processes whole documents instead of words, grouping them together in a vector space.

The Transformer (Vaswani et al., 2017) is a neural network architecture that enables parallel processing of input tokens instead of sequential processing. The architecture allows each input token to attend to all others with the use of a self-attention mechanism. By leveraging self-attention over the full context, the model is more capable of learning long-range dependencies, resulting in better disambiguation of polysemous words and understanding of nuanced linguistic structures. Transformers-based models such as BERT (Devlin et al., 2019), T5 (Raffel et al., 2020), and GPT (Brown et al., 2020) are now in standard use in all areas of NLP. A large portion of this thesis used Transformer-based models for vectorization due to their state-of-the-art performance in NLP tasks.

2.1.2 Dimension Reduction

Dimension reduction is used for reducing a high dimensional vector space to a lower dimension. This makes the data easier to handle (e.g., with limited

[1]Referring to the training mode Continuous Bag-of-Words. Skip-gram is another training mode that predicts the surrounding words based on the input word.

memory) and may improve its usefulness for downstream tasks.[2] Clustering algorithms have been reported to have challenges in high-dimensional vector spaces (Steinbach et al., 2004; Zimek et al., 2012). Thus, dimension reduction can be applied to reduce the negative effect of shrinking variance in distance between the vectors.

There are two main categories of dimensionality reduction techniques, those based on matrix factorization and those based on neighbor graphs. Principle Component Analysis (PCA, by (Pearson, 1901; Hotelling, 1933)), is a well-known and widely used example of matrix factorization. In combination with document embeddings it is common to use PCA to improve memory efficiency (Raunak et al., 2019). Two strengths of PCA are the computational efficiency, and the axes interpretability since they represent direction of greatest variance.

Neighbor graph methods are non-linear techniques that determine the relationships between data points before projecting them into a lower-dimensional space while preserving these relationships as much as possible. A widely used approach is t-Distributed Stochastic Neighbor Embedding (t-SNE (van der Maaten and Hinton, 2008)), which is commonly applied for visualization. A more recent method, Uniform Manifold Approximation and Projection (UMAP, McInnes et al. (2018)), is based on constructing a high-dimensional graph representation of the data and optimizing a lower-dimensional embedding to preserve its local and global structure.

UMAP was used throughout the work in the thesis as a standard component of the document clustering pipeline. The drawback for UMAP is similar to other neighbor graph methods in that the axes do not hold meaning. From (McInnes et al., 2018): "The dimensions of the UMAP embedding space have no specific meaning, unlike PCA where the dimensions are the directions of greatest variance in the source data". Still, it has been shown to improve the performance of subsequent clustering algorithms (Allaoui et al., 2020).

2.1.3 Clustering

Clustering is the tasks of grouping data points together based on their similarity. The clustering algorithms considered in this thesis are distance-based, meaning that they measure similarity between points in terms of distance in a vector space.

A popular partition clustering algorithm is K-Means (Lloyd, 1982). It initializes k centroids, assigns each data point to the nearest centroid, then updates the centroids based on the mean of assigned points, repeating until convergence. The algorithm is efficient but sensitive to initialization, assumes

[2]The term *curse of dimensionality* (Bellman et al., 1957) refers to the algorithmic challenges posed by high-dimensional vector spaces, where the volume of the search space increases exponentially with each added dimension. In clustering, this effect causes distances between points to converge toward the median distance, making it difficult to distinguish meaningful structures.

spherical clusters, and struggles with non-linearly separable data.[3] A common density-based clustering algorithm is Density-Based Spatial Clustering of Applications with Noise (DBSCAN, Ester et al. (1996)). DBSCAN locates clusters of points in the vector space depending on a search radius parameter ϵ. Thus, DBSCAN automatically identifies the number of clusters in the data and can identify outliers in low-density regions. A hierarchical extension of DBSCAN is HDBSCAN (Campello et al., 2013) which better adapts to density variations in the vector space. The distances to nearby points are calculated using the local density and then a minimum spanning tree is created over the points. From that tree, an hierarchy is created and then converted to a flat structure depending on a parameter *min_cluster_size*, which determines a lower bound on the size of clusters. HDBSCAN is used throughout the thesis and further explained in Section 3.1.

2.1.4 Evaluating Clusters

The clustering literature provide metrics for calculating the quality of clusterings. Standard clustering algorithm metrics can be divided into two groups. The use of external information in *extrinsic* metrics and only considering information present in the model as *intrinsic* metrics. Extrinsic metrics make use of labels to determine whether a model is accurate. Metrics such as Adjusted Rand Index (ARI, Hubert and Arabie (1985)), Adjusted Normalized Mutual Information (AMI, Vinh et al. (2010)), and F1-score fall into this category. Intrinsic metrics make use of information present in the model such as the cluster algorithm-assigned labels and vector space distances. They can be used to determine if an optimal clustering has been reached based on the layout of the vector space. Common metrics include the silhouette score (Rousseeuw, 1987), Davies-Bouldin index (Davies and Bouldin, 1979), and Calinski-Harabasz index (Caliński and Harabasz, 1974). The intrinsic metrics are often applied to measure how well-formed the clusters are depending on the vector space layout. The *compactness* of a cluster is how close within-cluster points are in relation to the whole dataset. *Separation* indicates how far the cluster is from its neighboring clusters.

The intrinsic metrics yield limited information if a well-formed vector space created by the vectorization step cannot be assumed. Additionally, the extrinsic metrics are not interesting unless the labels provided are clearly representative for the task. Using labels that are provided by public datasets or benchmarks is not suitable for evaluating clustering in specialized tasks. Therefore, it becomes necessary to collect human interpretation data of clusters specialized for a task, both to learn more about the vector spaces that the language models produce in detail, but also to improve clustering in a direction so that clusters are more interpretable for humans.

[3] See https://scikit-learn.org/stable/modules/clustering.html for an intuitive overview of different clustering algorithms.

2.2 Topic Modeling

Topic modeling is an unsupervised or semi-supervised machine learning technique for uncovering the core themes of a corpus. Topic modeling and document clustering have the similar motivation to assist in understanding a corpora. The main difference being that a topic model focuses primarily on explaining the corpus to a human by presenting coherent topics, while document clustering aims to structure the underlying documents to groups of similar contents. The *topic* in topic modeling has traditionally been defined as a probability distribution over a vocabulary. Often, the top k probable words in the vocabulary is presented to humans as the topic. Thus, the aim for a topic model has been to optimize models so that the top k words appear coherent to humans i.e. optimizing for topic coherence.

A comprehensive summary of the evolution of topic models can be found in (Churchill and Singh, 2022). The staple topic model Latent Dirichlet Allocation (LDA, Blei et al. (2003b)) was motivated by moving beyond Bag-of-words approaches and has essentially coined the family of models called topic models. In LDA, a topic is a probability distribution over a fixed vocabulary, and a document is a probability distribution over topics, both drawn from Dirichlet priors. Numerous extensions has been built on the standard model such as temporal (Blei and Lafferty, 2006; Wang and McCallum, 2006) or hierarchical (Blei et al., 2003a; Yee Whye Teh and Blei, 2006). With the introduction of the language model Word2Vec, authors also started incorporating contextual word embeddings for improving the topic models (Bianchi et al., 2021; Angelov, 2020). Neural topic models (Zhao et al., 2021) kept evolving with the Embedding Topic Model placing topic and document embeddings in the same vector space (Dieng et al., 2020). Also basic clustering of documents as described in Section 2.1.3 with algorithmic keyword extraction such as BERTopic (Grootendorst, 2022) were introduced as to be sufficient for creating coherent topics (Zhang et al., 2022; Sia et al., 2020).

In the intersection where topic modeling meet document clustering with language model embeddings, there exist a lack of human validation of the results. Model makers make claims about their models by consulting only automatic topic coherence metrics and not performing human evaluation. Here is where this thesis provides methods and addresses questions regarding the human interpretation of document clusters in order to reduce reliance on current automatic metrics.

2.2.1 Topic Model Evaluation

Successful topic models produces keyword sets representing the topics that are descriptive of the underlying data and interpretable to a human. How to define and measure coherency is a long standing problem for topic modeling practitioners (Doogan and Buntine, 2021). Lipton (2018), referring to model interpretability in machine learning: "The concept of interpretability

appears simultaneously important and slippery". In topic modeling, the human interpretation could either come from examining the keyword sets on a topic-level (Chang et al., 2009; Mimno et al., 2011; Newman et al., 2010) or on the document-level (Chang et al., 2009; Bhatia et al., 2017; Morstatter and Liu, 2017; Alokaili et al., 2019; Lund et al., 2019). Moreover, the collection of human interpretation data could come in a indirect form where evaluators perform a task, or direct form where the evaluators rate the cluster. The most generally accepted indirect data collection is the Intrusion task presented by (Chang et al., 2009), where crowdsource participants are given a keyword set with one *intruder* word which they are asked to identify. A high accuracy of the identification of the intruder intuitively means that the other keywords are conceptually coherent. An example of direct evaluation is having evaluators rate the topic representation (keywords or documents) on a scale of how coherent they are. It is widely considered that humans untrained for a task without feedback are poor at estimating quality (Morstatter and Liu, 2017; Doogan and Buntine, 2021). Thus, indirect measurements are considered better for a crowdsourcing environment, while direct measurements could be more suitable when the evaluators are trained domain experts.

Automatic topic coherence metrics are established when the human evaluation data are correlated with automatically calculated metrics in order to efficiently evaluate models without costly human evaluation. Some of the most common metrics being C_V (Röder et al., 2015), Normalized Pointwise Mutual Information C_{NPMI} (Lau et al., 2014; Aletras and Stevenson, 2013), and C_{UMASS} (Mimno et al., 2011). The topic coherence metrics are calculated using word co-occurrence which shares similar concepts to how neural language models operate when learning word contexts. The cost-saving potential of using these automatic coherence metrics has caused practitioners to rarely employ human validation of models, and simply use the results from automatic coherence evaluation for comparing models and guide parameter choices (Hoyle et al., 2021).

Criticism towards automatic topic coherence metrics stems from the results of human evaluation in specialized domains pointing to a weak correlation between the metrics (Doogan and Buntine, 2021). Hoyle et al. (2021) criticize that neural word representations have access to the same information as the coherence metrics such as NPMI, meaning that the topic models built on these representations can produce topics that score highly without being more interpretable. Hoyle et al. (2022) expand on this argument, meaning that the keyword-based paradigm is not stable enough for evaluating neural topic models. While some work suggests that current coherence metrics remain viable given an appropriate reference corpus (Lim and Lauw, 2023), there is broad consensus that automatic metrics will continue to play an important role in topic modeling research. Interesting avenues forward suggest the using LLM to essentially replace the crowdsource workers performing the evaluation tasks (Rahimi et al., 2024; Stammbach et al., 2023).

The evaluation framework presented in this thesis is foremost for evaluating

document clustering, but it also gives a fresh angle for discussing human interpretation in topic modeling evaluation. Additionally, the framework is designed to be flexible to address the need for a standardized method for practitioners to validate their topic models in specialized fields.

Chapter 3

Methodological Overview

In this chapter, the overview of the datasets and the clustering pipeline are presented. Additionally, an overview is given of the novel tools for evaluating document clusters through human interpretation, STELLAR and CIPHE.

3.1 The Clustering Pipeline

The clustering pipeline presented here was used throughout Papers I-IV for creating document clusters. It is a straightforward combination of components for vectorization, dimension reduction and clustering. It is popular for its versatility and modularity, and the same pipeline used in BERTopic (Section 3.1.1). The pipeline is found in many works of documents clustering with varying components (Curiskis et al., 2020; Radu et al., 2020).

Language models The language models used were mainly the Transformer-based models BERT (Devlin et al., 2019) and Sentence-T5 (Ni et al., 2022). The Transformer-based models results in 768D vector embeddings of the text. Generally, English base models were used for vectorizing without further training.

Dimension reduction The dimension reduction component UMAP is the default component in BERTopic and also used throughout the work in this thesis. UMAP has demonstrated improved performance in clustering tasks (Allaoui et al., 2020), and works well with visualization. UMAP first builds a k-nearest neighbor graph, connecting each data point to its closest neighbors. Then, using fuzzy set theory, it estimates the overall topology of the underlying manifold (the global shape) of the data in high-dimensional space. Finally, it optimizes a low-dimensional representation to preserve the structure of the high-dimensional space as closely as possible. UMAP can be thought of as performing a form of partial clustering on the vector space, as it leverages local

neighborhood density to define the topology. This process of connecting points based on local density may enhance clustering performance in downstream tasks as there is a tendency for creation of "hubs" in high-dimensional data that can be exploited using k-neighbors (Tomasev et al., 2014).

Clustering The main clustering algorithm used in Papers I-IV was the HDBSCAN algorithm. The benefits of HDBSCAN in comparison to DBSCAN is that it is less sensitive to differences in density in the vector space and that it requires no preset number of clusters which is advantageous for discovering what topics exist in the corpus (Campello et al., 2013). HDBSCAN first creates a graph network between points that takes into account the local density of points. Then, the graph is structured for clustering by removing weak links and creating a hierarchy of the network nodes. Lastly, the algorithm chooses the most stable groups to become clusters based on the parameter *min_cluster_size*, automatically determining how many clusters to create and labeling all points not in a stable group as noise.

3.1.1 BERTopic

BERTopic (Grootendorst, 2022) is a topic model built on the described clustering pipeline. The standard components are vectorization with a Sentence-BERT (Reimers and Gurevych, 2019), UMAP for dimensionality reduction, and HDBSCAN for density-based clustering. BERTopic becomes a topic model by incorporating the keyword extraction method Cluster-Term Frequency-Inverse Document Frequency (C-TF-IDF) to create a topic representation for the clusters. C-TF-IDF operates similarly to standard TF-IDF (Section 2.1.1), but instead of individual documents, it treats all text within a cluster as a single document. The inverse document frequency is then calculated using the concatenated text from other clusters. The incorporation of keywords in BERTopic made it possible to compare the results from clustering evaluations with topic coherence metrics.

3.2 Datasets

A key focus of this thesis was the application of the clustering pipeline in a real-world industry setting. To evaluate the clustering pipeline and evaluation methods in a real application environment, all papers included unique industry news datasets provided by Aeterna Labs. In addition to the scraped industry datasets, other public news datasets and documents from other domains were included. A summary of the datasets and in which paper they are used can be found in Table 3.1.

Dataset	Abbreviation	Documents	Time period	Region	Paper
Scraped UK News	UKNEWS	10 000	220529–220622	UK	I, II
Scraped UK News 1	SUN1	27 786	240101–240813	UK	III
Scraped UK News 2	SUN2	22 937	241014–241027	UK	III
Scraped News Articles Classified w. Keywords	SNACK	16 391	2021	UK	IV
Scraped News Articles Multilingual	MULTINEWS	30 000	2024	SWE, DEN, UK	V
Wikipedia Current Events 2018	WCEP18	59 073	2018	Global	II
Wikipedia Biographies	WIKI	50 000	fetched 230424	Global	III
Yelp Reviews	YELP	50 000	050504–220119	Mainly USA	III
AG News	AG News	60 000	2004–2020	Global	IV
Thomson Reuters Text Research	TRC2	11 085	080101–081212	Mainly USA	IV

Table 3.1: Datasets used in the thesis. Every paper used an unique industry dataset provided by Aeterna Labs.

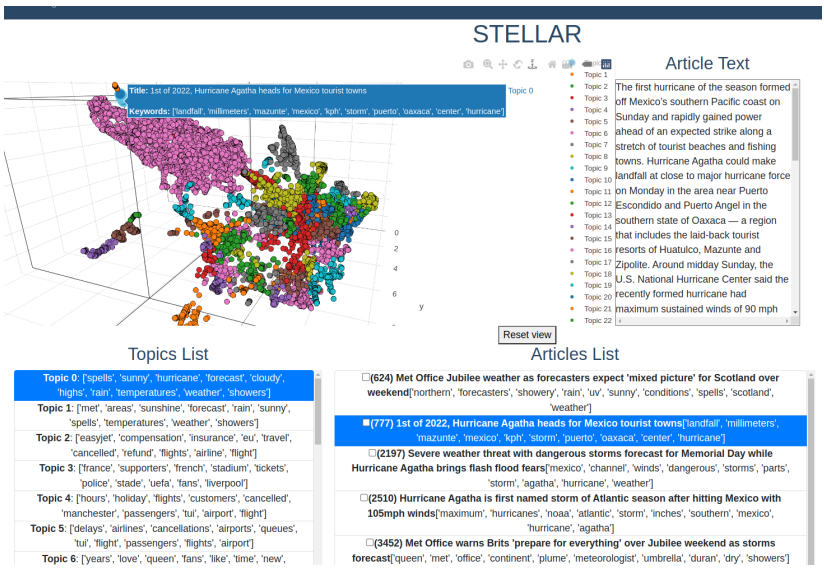


Figure 3.1: The topic browser STELLAR. From Paper I (Eklund and Forsman, 2022).

3.3 Cluster Sample Validation Process

Central to the thesis is the qualitative human evaluation of clusters here defined as *cluster sample validation*. Consider having made a clustering of documents. A natural validation method that practitioners may perform is to randomly sample documents from a cluster and inspect them. Usually, a coherent cluster should have an general pattern that is easily identifiable by the evaluator, and it should be possible to give the cluster a summarizing name. Additionally, the evaluator may look for documents that are not fitting with the rest of the cluster to determine if the clustering model has made any errors. The systematic process for inspecting multiple clusters in this way is here defined as the *cluster sample validation process*.

The two developed tools STELLAR and CIPHE were based on the cluster sample validation process. STELLAR (Paper I) is a topic browser that focuses on visualizing the whole clustering through a 3D-visualization, compact lists of documents, and cluster summaries to aid exploration of the data. CIPHE (Paper II & III) is a structured framework built around the process that can be used as a survey tool to gather information on cluster precision and characteristics. The overview of the two tools is presented in the following sections.

3.3.1 STELLAR

Systematic Topic Evaluation Leveraging Lists of ARticles (STELLAR, Figure 3.1) is a tool introduced in Paper I for visualizing clusters and their contents in a compact view. The idea behind STELLAR was to allow for a user to easily access the documents that belong to a cluster which enables efficient exploration of the corpus. STELLAR consists of three main components: 1) An interactive 3D-visualization of the clusters, 2) a list of clusters with their description, and 3) a list of documents for each cluster. The raw document text is also viewable in a window.

As Paper I described, the strength of STELLAR lies in it being a topic browser that allows for in-depth exploration of clusters beyond the top keywords. It was designed with the considerations of an expert evaluator in mind, which limited its capabilities as a framework for crowdsource participants. Thus, new work began for creating a framework more fit for structured evaluation that requires less training of the evaluators.

3.3.2 CIPHE

Cluster Interpretation and Precision from Human Exploration (CIPHE) consists of a survey platform and metrics for calculating cluster precision and characteristics. The main ideas from the cluster sample validation process were converted to a framework for collecting human interpretations of samples of clusters.

Tasks The framework consists of three tasks for the evaluator to perform in a survey platform that can be seen in Figure 3.2. The tasks are:

1. **Inclusion task:** The participant is asked to explore the cluster sample by reading the titles and navigating through the text bodies. The participant is prompted to decide which documents, according to them, do not belong to the cluster.

2. **Naming task:** The participant is asked to name the cluster using their own words.

3. **Likert assessment task:** The participant is asked to answer Likert-scale questions (Schuff et al., 2023; Joshi et al., 2015) about the cluster. The Likert statements concern the perceived difficulty of performing the inclusion and naming tasks, and aspects particular to the specific evaluation at hand. The Likert statements enable asking about characteristics that the investigator has defined, such as negative emotional response, which makes the framework adaptable to specialized needs.

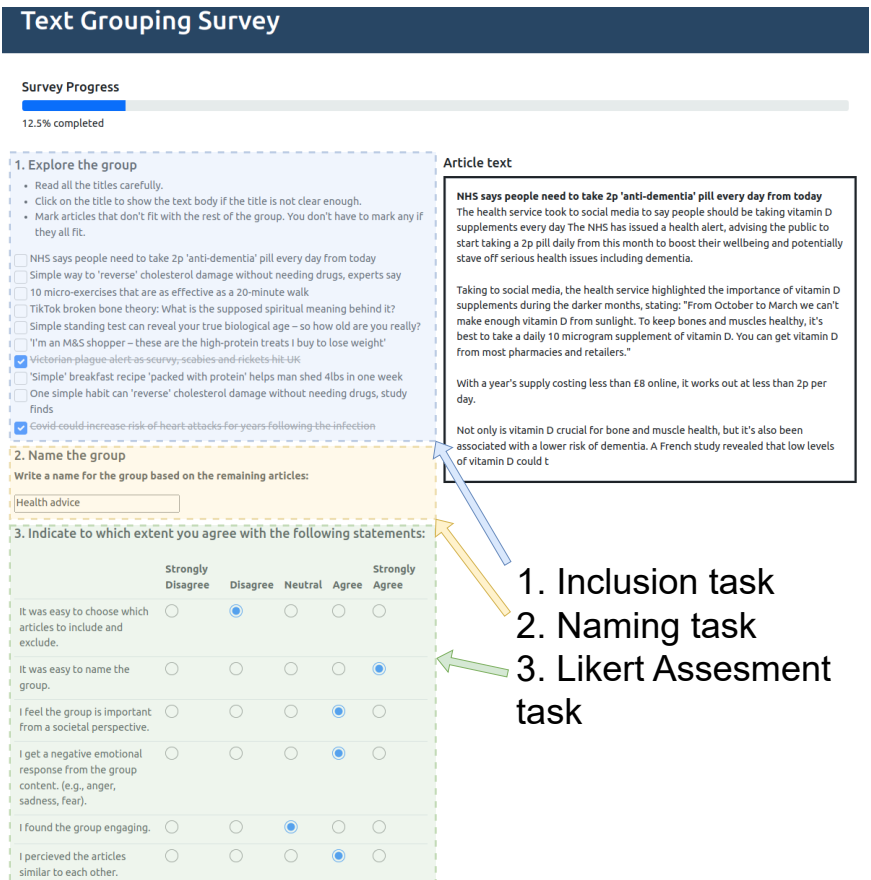


Figure 3.2: The CIPHE platform with the three evaluator tasks highlighted. 1) The Inclusion task where the evaluator explores the cluster sample and removes articles not belonging there. 2) The Naming task where the evaluator names the cluster in their own words. 3) The Likert Assessment task where the participant give their interpretation of different cluster characteristics. (Eklund et al., 2025a)

Metrics The data collected from these tasks are structured with the CIPHE metrics. The metrics concerns individual clusters to allow for cluster-by-cluster comparisons which also lends them for comparison with clusters from other models. The metrics are computed by taking the averages over the answers of all participants having evaluated a particular cluster as follows:

- **Cluster precision** (CP): The averaged number of included documents (in the inclusion task) divided by the total number of documents in the sample gives a precision estimation of the cluster.
- **Inclusion agreement** (A^{inc}): The average agreement on the decision to include individual documents in the sample.
- **Naming agreement** (A^{name}): The average cosine similarity between the evaluator-given names in the naming task.
- **Likert assessment** (L): The average score of the Likert scale {strongly disagree, disagree, neutral, agree, strongly agree} converted to their respective numerical value $0, 0.25, 0.5, 0.75, 1$.

The CP metric is intended to give a score that can be used for algorithmic improvement. The other metrics gives a more nuanced view of the cluster, beyond what is possible with a single precision score. The inclusion and naming agreement are intended to indicate whether the evaluators are interpreting the sample in a similar way. The inclusion agreement gets a lower score if the evaluators are choosing different documents to include. The naming agreement gets a lower score when participants are writing semantically dissimilar names for describing the cluster. The Likert assessment records interpretation from other perspectives than CP. The evaluators are asked to rate the difficulty in performing the inclusion and naming tasks. Additionally, the Likert scale questions allow the recording of other (for the investigator) interesting characteristics, such as evaluator opinion on perceived impact on society for a certain news story. For precise definitions of the metrics used in CIPHE, see Paper III.

Evaluation sets and sample size The standard number of documents from each cluster that was shown to an evaluator was 10. However, when using CIPHE in a real setting, it is not sufficient to estimate cluster precision and characteristics with only a single sample of 10 documents. Thus, the concept of evaluation sets was introduced.

To create an evaluation set, a sample size s is determined by using Sampling from a Finite Population (Körner and Wahlgren, 2015) which depends on the cluster size N, estimated proportion of included documents π, and allowed margin of error ϵ. A sample of s documents is sampled from the cluster and divided into evaluation sets of size m each. In cases s was not divisible by m, more documents were sampled. E.g. given a cluster with $N = 1000$ documents the required sample was determined to $s = 88$. Then, 9 unique evaluation

sets of size $m = 10$ each were created after sampling two more documents. The standard evaluation set size for CIPHE has been $m = 10$ throughout this thesis but can be freely adjusted.

Chapter 4

Summary of Papers and Contributions

The main scientific contribution of the thesis is the development of methods for collecting human interpretation data, which enables new approaches for evaluating document clustering through human interpretation. In Paper I, the cluster sample validation process was first presented in an industry validation study with the STELLAR tool. Building upon the ideas of Paper I, the CIPHE framework was established in Paper II together with a comparison of participant instruction sets. In Paper III, CIPHE was improved to its current state and used for studying limitations in topic modeling coherence evaluation. Paper III is the most recent work and a longer article summarizing the work with CIPHE, thus making it the most important contribution. Paper IV was a structured parameter study of the document clustering pipeline, with the main aim to investigate the effect of the dimension reduction component. Finally, Paper V aimed to reduce the evaluation cost when performing continuous human-in-the-loop classifier validation of complex systems in their application environment, an aim that was directly derived from the real industry setting of the thesis.

Author contributions The author of the thesis was the main contributor for all the papers included in the thesis. Specifically, idea conception and development, literature review, experiment design, code design, and authoring papers. Throughout the studies, close collaboration was conducted with Frank Drewes and Mona Forsman for idea development, discussion and authoring papers.

4.1 Paper I

Topic Modeling by Clustering Language Model Embeddings: Human Validation on an Industry Dataset (Eklund and Forsman, 2022)

> *"The paper that introduces STELLAR and first exemplifies the cluster sample validation process."*

Paper I was a validation study of the clustering pipeline on an unique industry dataset provided by Aeterna Labs. The paper addressed the lack of human validation of neural topic models based on the clustering pipeline (Grootendorst, 2022; Sia et al., 2020). STELLAR was introduced as a topic browser paired with an evaluation protocol based on the cluster sample validation process. The expert human evaluators were provided 20 articles from each of the 63 clusters. Then, they used the introduced STELLAR to explore the clusters and mark articles that they considered not coherent with the rest of the cluster. The cluster coherence was calculated in the same manner as CIPHE calculates cluster precision.

The results showed that 52 out of the 63 clusters produced by the topic model were deemed coherent for this news dataset. However, 43% of the articles ended up in incoherent clusters. Particularly, the largest cluster was deemed incoherent which meant that a large portion of the total set of articles were in incoherent clusters. Further analysis attributed this to be the effect of issues in the vectorization process, along with the soft clustering approach forcing outlier points into clusters, leading to higher article misplacement.

Moreover, limitations with the newly introduced evaluation process was highlighted and discussed. Particularly, it was evident that a reduction in coherence often came from evaluators settling for different granularity levels (see Table 1.1 in the Introduction). The paper also concluded that incoherent clusters were often a mix of themes which highlighted a possible improvement area for the clustering model.

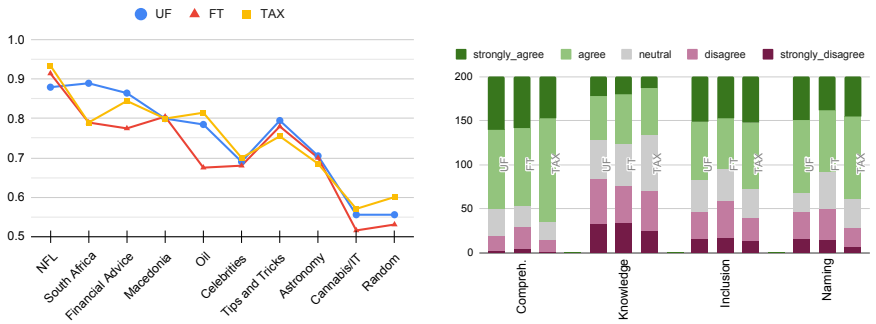
Paper I is important for the thesis as it was a first attempt to provide human evaluation of topic coherence without the use of keywords. The insights gained through this project ultimately led to the development of the CIPHE framework.

4.2 Paper II

CIPHE: A Framework for Document Cluster Interpretation and Precision from Human Exploration (Eklund et al., 2024)

> *"The paper that introduces CIPHE and compares instruction sets for collecting human interpretation."*

Paper II addressed limitations to the method in Paper I and introduced CIPHE as a crowdsource framework for collecting human interpretation of clusters. The participants took the CIPHE survey with three different instruction



(a) Comparison of CP, clusters sorted by C_V.

(b) Comparison of Likert responses.

Figure 4.1: Results from Paper II indicating that the instruction sets yield similar responses from the participants. (Eklund et al., 2024)

sets of varying degrees of freedom to name the clusters. Those were a) free text naming (FT), b) lightly guided to choose a unifying meta-category before naming (Unifying Feature (UF)), and c) choosing a category from a taxonomy before naming (TAX). The aim was to examine CIPHE in a crowdsource environment and to gain knowledge about how to set up the instructions in a CIPHE survey.

The results (Figure 4.1) showed that varying the instructions had little impact to the outcome of the evaluation, meaning that participants mainly assessed the clusters based on their interpretation of the articles and were not largely affected by the provided guidance in the instructions. This meant that the instruction for free-text naming was considered sufficient for most tasks.

CIPHE introduced the inclusion task together with the CP metric for estimating model misplacement of documents. By comparing the results from the CP metric and the expected values based on our (the authors) opinion, there were signs that the inclusion task, which produces the precision score, had the potential to function as a novel indirect human interpretation data collection task, filling the same function as the intrusion tasks by Chang et al. (2009). Further experimentation with CIPHE was needed to confirm such a claim, and to make a comparison with the family of methods in topic coherence.

4.3 Paper III

Comparing Human-Perceived Cluster Characteristics through the Lens of CIPHE: Measuring Coherence beyond Keywords (Eklund et al., 2025a)

> *"The paper that improves the ability of CIPHE to characterize clusters and that compares the framework with keyword-based methods for topic modeling evaluation."*

The limited scope in Paper II motivated a more thorough study for investigating what types of conclusions could be drawn about cluster models with the use of CIPHE compared to traditional keyword-based methods (KWM) often found in topic coherence evaluation. In Paper III, a study with 122 participants was conducted where KWM for collecting topic coherence was examined through CIPHE. Human interpretation data was collected for 21 clusters from three different datasets and analyzed in-depth. Additionally, a case study to investigate the capabilities of CIPHE when applied to a full corpus was needed. An experiment was conducted with 180 participants evaluating 37 clusters from the UK news domain created with the clustering pipeline.

The results from the KWM-CIPHE comparison showed that there are limitations to how keywords represents clusters, both for the precision score and for other characteristics. The resulting precision score from the word-intrusion task (Chang et al., 2009) fluctuates a lot compared to the resulting CIPHE cluster precision from the inclusion task. The fluctuations are attributed to the number of keywords and the choice of intruders shown to the participant. One of the strengths of CIPHE is that it makes such engineering with intruders unnecessary. Intruders in CIPHE is further discussed in Section 5.1.

The conclusion was that keyword representations have limited capabilities of describing characteristics other than a main theme. When the documents in a cluster has some unexpected underlying sentiment, a keyword representation will struggle to convey it. As an example, a cluster contained news about a canceled Taylor Swift concert due to a terror threat. The keywords ⟨`eras`, `tickets`, `wembley`, `stadium`, `fans`, `tour`, `concerts`, `swift`, `taylor`, `vienna`⟩ indicated an entertainment cluster, so participants exposed to KWM used their expectation of an entertainment cluster to characterize it. This characterization was drastically different to the participants exposed to CIPHE, who could easily identify the darker theme. Clusters with an unexpected underlying sentiment may be rare but could have significant impact when using the models for a specific purpose such as contextual advertising.

The case study demonstrated how CIPHE can be used for a research purpose similar to Paper I but with crowdsource participants. Qualitatively, the characteristics determined by CIPHE are in line with what one may intuitively expect (Table 4.1) which showed the viability of using CIPHE to collect and structure human interpretation data. Clusters ranked by the CP metric also

indicated that broader clusters were less coherent (Figure 4.2) which gives informative input for cluster pipeline improvements. Still, limitations remain, such as the evaluator not having access to an overview of the full cluster context, or that the instructions are misinterpreted in a crowdsource environment. These limitations are further discussed in Section 5.1.

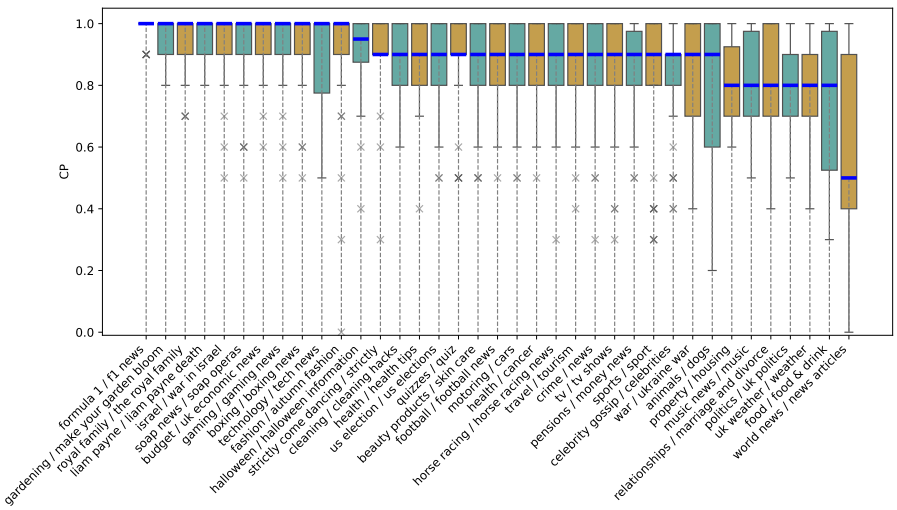


Figure 4.2: The CP scores for the different clusters in the UK case study in Paper III. The clusters are sorted by their median score (marked in blue), which indicates that clusters with a broader theme tends to get a slightly lower score while narrower themes such as events get a higher score. (Eklund et al., 2025a)

	Negative Emotion		Impact		Engagement	
	Cluster	Score	Cluster	Score	Cluster	Score
Top	*Russia-Ukraine war*	0.68	*Pensions & benefits*	0.83	*Russia-Ukraine war*	0.67
	Crime	0.67	*UK economic news*	0.83	*Israel-Hamas war*	0.66
	Israel-Hamas war	0.66	*Israel-Hamas war*	0.81	*Food & drink*	0.66
	Liam Payne passing	0.60	*Russia-Ukraine war*	0.79	*UK economic news*	0.63
	Health & hospital	0.59	*Health & hospital*	0.77	*Health & hospital*	0.62
Bottom	*Quizzes*	0.20	*Boxing*	0.36	*Boxing*	0.42
	Formula 1	0.19	*Gaming*	0.35	*Royal family*	0.42
	Halloween	0.18	*Horse racing*	0.34	*Strictly come dancing (TV)*	0.39
	Food & drink	0.18	*Soap operas (TV)*	0.33	*Soap operas (TV)*	0.37
	Gardening	0.07	*Celebrity gossip*	0.29	*Celebrity gossip*	0.34

Table 4.1: Ranked comparison of clusters with respect to the Likert assessment scores for the characteristics Engagement, Impact, and Negative Emotion. The top and bottom five for each characteristic are shown. (Eklund et al., 2025a)

4.4 Paper IV

An Empirical Configuration Study of a Common Document Clustering Pipeline (Eklund et al., 2023)

> *"The paper that examines the BERTopic pipeline for clustering labeled datasets with a focus on dimension reduction."*

The document clustering pipeline described in Section 3.1 was examined through a large parameter search on three labeled news datasets. Two algorithms for each component of the pipeline were compared. Vectorizing (Doc2Vec and BERT), dimension reduction (PCA and UMAP), and clustering (K-Means and HDBSCAN).

The conclusion from the study was that the vectorization component holds the most influence on the resulting clustering, and that for the two vectorization models tested in the study, BERT is to be preferred over Doc2Vec. While the conclusion may appear intuitive in retrospect, at the time of our study, there was genuine uncertainty regarding whether the increased complexity of the Transformer architecture and the extended training context of BERT would improve document clustering performance. The hypothesis was that Doc2Vec may disregard finer details in the texts which would lead to the capturing of broader semantic themes in the vectors that would facilitate clustering. However, the empirical results (Table 4.2) showed that BERT performed on-par with or better than Doc2Vec in all comparisons, which consequently lead to the recommendation of using Transformer-based models with the clustering pipeline.

Regarding the dimension reduction, the results provided empirical support for the hypothesis that the neighbor-based dimension reduction of UMAP has advantages also in the text domain (Allaoui et al., 2020). Our results also indicated that clustering performance dropped when the dimensionality was reduced to lower than 15D. Thus, this paper provides empirical support for using UMAP, and not reducing the dimension to lower than 15D.

Respecting the often specialized applications of news clustering and that the study only used 3 labeled datasets, general recommendations to parameter settings could not be given from the result of this study. This paper highlights the challenges associated with relying on benchmark datasets for evaluating clustering algorithms. The labels in the datasets came from broad news categories, tags from the publishers in the Reuters news dataset, and classes from one validated classification model from Aeterna. Whatever parameters found in this study only indicates what would be recommended for a study in closely similar setting and labeling as the datasets in the study. In a real-world application of clustering, adaptable evaluation methods are needed. This was the first project in the doctoral studies, which lead to the pursuit of efficient and adaptable evaluation methods.

Data	Setup	Dim	Alg. setting	Nr of clusters	Time	*AMI*
SNACK	bert_umap_kmeans	25	6	6	10.81	0.54
	bert_umap_hdbscan	15	320	7	20.27	**0.55**
	doc2vec_umap_kmeans	15	6	6	11.33	0.54
	doc2vec_umap_hdbscan	25	640	6	16.55	**0.55**
	bert_pca_kmeans	25	6	6	2.62	0.51
	bert_pca_hdbscan	15	160	6	6.37	0.49
	doc2vec_pca_kmeans	50	6	6	2.02	0.49
	doc2vec_pca_hdbscan	7	160	6	3.23	0.45
AG NEWS	**bert_umap_kmeans**	10	4	4	31.06	**0.64**
	bert_umap_hdbscan	25	2560	3	68.6	0.63
	doc2vec_umap_kmeans	3	8	8	30.26	0.31
	doc2vec_umap_hdbscan	15	160	5	75.11	0.31
	bert_pca_kmeans	50	4	4	5.82	0.6
	bert_pca_hdbscan	5	2560	4	16.79	0.54
	doc2vec_pca_kmeans	50	16	16	11.09	0.24
	doc2vec_pca_hdbscan	3	80	5	4.29	0.15
REUTERS	bert_umap_kmeans	50	7	7	14.62	0.69
	bert_umap_hdbscan	25	160	10	14.47	**0.71**
	doc2vec_umap_kmeans	10	56	56	8.1	0.24
	doc2vec_umap_hdbscan	10	10	62	8.1	0.24
	bert_pca_kmeans	25	14	14	2.18	0.6
	bert_pca_hdbscan	15	20	16	2.33	0.66
	doc2vec_pca_kmeans	50	224	224	10.27	0.24
	doc2vec_pca_hdbscan	25	5	47	6.29	0.18

Table 4.2: From Paper IV (Eklund et al., 2023), a table of the best cluster pipeline configuration according to the extrinsic metric AMI for each dataset. The column 'Alg. setting' reports the number k of clusters in K-Means and *min_cluster_size* in HDBSCAN.

4.5 Paper V

Industry Quality Control for Efficient Continuous Human Validation of Deployed Text Classification Systems (Eklund et al., 2025b)

> *"The paper that presents a human-in-the-loop algorithm, along with Acceptance Criteria, for validating classifiers in a production environment."*

When working with the inclusion task of CIPHE, it became apparent that methods for assessing texts depending on some investigator-decided criteria has wider industry application. Paper V presented an adaption of production industry concepts of Quality Assurance, in particular Acceptance Sampling, to the NLP domain. The Integrated Acceptance Sampling algorithm is suitable when testing costs are high such as in human evaluation of text classification. The most notable difference between production industry and NLP is in creating assessment criteria for how products are evaluated contra text. E.g. it

	English	Swedish	Danish
Accuracy	0.919 ± 0.027	0.914 ± 0.028	0.872 ± 0.034
α	0.82	0.77	0.80
Consensus	97	93	93

Table 4.3: The classification accuracy of Llama2-13B estimated with human evaluation on a sample of 370 articles per language. The inter-rater agreement was calculated with Krippendorff's α on 100 articles. Consensus is the number of pairs where all three evaluators agreed. (Paper V (Eklund et al., 2025b)).

is easily assessed if an object breaks under a certain pressure while assessing if a text belongs to a certain category leaves much to the interpretation of the evaluator. The paper revolved around addressing this problem and introduced the concept of Acceptance Criteria for assessing document-class pairs.

The Acceptance Criteria builds on two main considerations. First, the *granularity* of a theme should be determined. E.g. an article about Olympic pole vaulter Armand Duplantis to be classified as Sports, Olympics, or Pole Vault. Second, the *strictness* depends on whether an article should exclusively discuss a certain theme, be mainly about a theme (but is allowed others), or just mention the theme. E.g. whether it is accepted to classify a year-in-review article as Pole Vault if Duplantis is mentioned. Together with other practical considerations of the classifier application environment, the investigator creates statements about the document-class pair that can either be True or False. As an example with an Olympics classifier, a statement could be *"The article exclusively discuss results, participants, and events at the Olympics"*.

In the experiment, a quantized Llama2-13B model was used to classify news articles in English, Swedish and Danish into 18 predetermined news categories. Then, the article-classification pair was assessed by three expert human evaluators according to pre-determined acceptance criteria created for the industry setting of contextual advertising with Aeterna.

The results from simulation showcased how the algorithm can save effort, especially identifying defunct classifiers quickly. However, the analysis of the human evaluation again showed difficulties in aligning views between evaluators (Table 4.3), with evaluator experiences having larger influence when the instructions leave room for interpretation. E.g. what entails the class *Public Sector* was not aligned between the evaluators. Considering that evaluation complexity is increasing due to rising complexities of modern NLP systems using LLM only further emphasizes that human validation is necessary. The presented human-in-the-loop validation algorithm Integrated Acceptance Sampling reduces the complexity to only validating the input-output pairs. But the study concluded that significant work should be put into the Acceptance Criteria to have meaningful validation of a specialized task. The work with the Acceptance Criteria statements laid the foundation for formulating the Likert assessment task in CIPHE.

Chapter 5

Synthesis and Discussion

The papers included in this thesis are contributions to evaluation methods within document clustering where human evaluation is needed to validate quality. In the following sections, the research aims are addressed through discussing the results from the papers and synthesizing with existing literature. First, the work with building STELLAR and CIPHE on the cluster sample validation process is evaluated. Next, the results with human interpretation of clusters is put in relation with topic modeling evaluation and topic coherence. Then, the clustering pipeline is discussed through the empirical results from the papers, in particular for news articles. Finally, potential future work is summarized with framework-specific improvements and the expectations for how CIPHE can be used.

5.1 Evaluating the Cluster Sample Validation Process

This thesis lays the groundwork for creating sophisticated evaluation methods that are adapted to specialized problem domains. Both the presented tools STELLAR and CIPHE build on the cluster sample validation process developed in this thesis. In this section, the development process, the current state of CIPHE, and its limitations are analyzed.

Paper I presented STELLAR and an expert evaluation for a unique news dataset built on the cluster sample validation process. The approach with the topic browser STELLAR worked adequately for well-instructed experts, but needed statistical refinements and protocol improvements for it to become a standardized evaluation framework. Paper II presents the standardized method with the introduction of CIPHE. The experiments conducted during the development of CIPHE address functionality in a crowdsource environment. Mainly, how the instructions should be formed in order to allow evaluators flexibility to respond using their own judgment, while retaining a similar understand-

ing among each other of what the task expects from them. In Paper III, the characterization beyond the main theme was central. The comparison with keyword-based methods showed how keyword-based representations may struggle to represent characteristics other than the theme, and how CIPHE could yield valuable insights when used to evaluate a clustering of the UK news domain.

The concepts of characterising clusters beyond theme in Paper III was an improvement based on the findings in Paper II. The first version of CIPHE contained an instruction set called Unified Feature, where the participants were expected to look beyond the general theme of the cluster and conclude that the documents shared a stylistic feature that unified them. E.g. a cluster containing articles on advice about money and finance, the participants were expected to notice the *Advice* feature. However, a majority of participants still answered thematically such as *Money*, *Pensions*, *Government Benefits*. The results of Paper II also failed to identify a real disadvantage to letting the participants freely name the cluster. Thus, the question about identifying the overall theme became the Naming task, and the Likert assessment task was extended to allow the identification of other characteristics. The updated framework successfully collects data on the characteristics of clusters which resonates with intuition as shown in Paper III.

A problem with the current state of CIPHE (mainly discussed in Paper II) is that evaluators seem inclined to remove at least one article in the Inclusion task. This creates a phenomenon where very coherent clusters may suffer a reduction in the precision score since crowdsource evaluators are not comfortable to include all articles. This was not a problem in Paper I where the expert evaluators had a more complete view of the corpus context and a clear understanding of the evaluation goal. To mitigate this effect, the idea of incorporating concepts from Intrusion (Chang et al., 2009) was discussed in Paper II. In that experiment, an example cluster contained 8 out of 10 articles clearly focused on one theme, while the remaining 2 were on another. Nearly all crowdsource evaluators could agree on what articles should be removed, which gave the cluster an undisputed score of CP = 0.8. A way to emulate this effect would be to incorporate the concept of intrusion to CIPHE by adding intruder documents to the evaluation sets. That way, the precision score is likely to have high agreement and be more accurate. Mainly opposing this idea was that gamifying the inclusion task may incentivize participants to focus on finding intruders and be less engaged in exploring the clusters through their own judgment. Moreover, the fluctuating results of MP in Paper III showed that setting up a survey with intrusion requires meticulous engineering to show "good" intruders to a participants. Finding intruder documents for CIPHE adds an unwanted layer of complexity where the intruding document will contribute to the overall context. This furthers the argument against the use of intruders in CIPHE. However, it remains a future research direction to provide more information on this matter.

If we isolate CIPHE as a framework and do not consider CIPHE research usecases, there are numerous aspects about CIPHE that can be further ex-

amined through experimentation. One such experiment could be comparisons of the evaluation set size m to determine an optimal number of documents per cluster viewed by the evaluator. Currently, the size is $m = 10$ to avoid participant fatigue, but the number does not have any experimental backing.

Another area of refinement is the statistical foundation of both CIPHE and the Integrated Acceptance Sampling algorithm in Paper V. The use of Sampling from a Finite Population provides a well-defined sample size with desired statistical guarantees. Yet, further analysis is needed to clarify its implications for CIPHE, in particular to the CP metric. A deeper statistical examination could strengthen the theoretical base of CIPHE, which would make the results more actionable and possibly improve efficiency in practical application.

A common problem for evaluators using CIPHE is not having access to an overview of the corpus context. It is common in other topic modeling evaluation framework to have some type of visual representation of the topics (Peirson et al., 2016; Grootendorst, 2022; Kardos, 2023). Such visualization was given in STELLAR in the form of a 3D projection and a list of all topics. That helped the evaluators know what granularity level was expected when performing the evaluation. However, it was lacking in CIPHE in favor of giving an explanation of the corpus in the instructions. This may be the reason for a lot of granularity-related issues in CIPHE. An experiment designed to determine the impact of giving the participants a cluster overview could be important for further development of the CIPHE framework.

The framework is new and the tasks and metrics are expected to evolve, either through general framework improvements or optimization for specialized tasks. The work in this thesis laid the foundation and showed the feasibility of converting the cluster sample validation process into a framework for document cluster evaluation.

5.2 Topic Coherence Examined through CIPHE

The structuring of documents into clusters of similar content generates valuable new information from large data collections. This information is critical not only for clustering experts but also for professionals across diverse fields, from medicine to advertising. For these clusters to be actionable, they must appear coherent to human evaluators. While the concept of coherence is central to clustering and topic modeling, its definition remains ambiguous. Most research in this area stems from topic modeling, where coherence is often evaluated through sets of keywords. Thus, this thesis has had multiple touch points with the research on topic coherence, which is discussed here.

In Paper III, the human-perceived coherence of clusters through CIPHE was compared with those using traditional keyword-based methods (KWM). This comparison revealed significant differences in how the clusters were characterized, and in particular how KWM struggled to represent subtle influences

on contexts. Participants exposed to keywords needed to use their experience to extrapolate from just a few words what cluster characteristics may be. This was required to a lesser extent for participants reading the raw texts in CIPHE. One example was the Taylor Swift cluster, where the keyword representation indicated that the central theme of the cluster was *Entertainment* or *Concerts*, but failed to capture the negative tone of a terror threat. These findings align with criticisms in prior research, which argue that keywords are insufficient for describing the complexities of human nature (Doogan and Buntine, 2021; Hoyle et al., 2021). Still, the paper concluded that keyword representations are likely to be sufficient for capturing the general theme of the cluster, which may be adequate for some research purposes applying topic or clustering models.

The automatic topic coherence metrics were also briefly examined in Paper III. The results from the correlation calculations between CIPHE, C_V and *NPMI* indicated moderate correlation. Instead, the most correlated metric candidate was the cluster metric Silhouette coefficient. This means that a new angle to automatically estimate topic coherence without relying on word co-occurrence could be to apply distance-based metrics from the clustering literature. This path heavily relies on the vectorization process reliably correlating with human perception of coherence, which is further discussed in Section 5.3.

An interesting way forward is using LLM to replace crowdsource workers performing the coherence evaluation (Rahimi et al., 2024; Stammbach et al., 2023). This has potential to be valuable in the model development phase, where CIPHE also contributes with novel tasks that could be performed by an LLM. Still, emphasis should be put on the use of human validation when making claims about human interpretation in research. If not with CIPHE, then at least the LLM performing the evaluation should be quality assured through a protocol similar to the one presented in Paper V.

5.3 Evaluating the Clustering Pipeline

The original reason to develop methods for evaluation was to validate clustering systems with the end goal of being deployed to a production environment in industry. The industry setting of contextual advertising required the system to have semantically coherent clusters to avoid placing advertisement in a context that is harmful to the advertiser. The use of the clustering pipeline in Papers I-IV — and many other projects not reported in the thesis — has collectively contributed to the knowledge about this particular setup. In this section, the clustering pipeline is discussed, in particular for news clustering.

Before discussing the clustering system, it is important to acknowledge certain limitations of this work. The majority of experiments were conducted using language models primarily trained on English. Unpublished trials with multilingual models, such as Sentence-T5, did not produce clusters to the same level of quality for e.g. Swedish and Danish. Additionally, the news corpora used in this research typically ranged between 10,000 and 60,000 articles. Scaling be-

yond these volumes is likely to require adjustments to parameters in UMAP and HDBSCAN. However, despite these constraints, the insights gained about the vectorization process should remain applicable to larger datasets and broader contexts.

The empirical results from this thesis suggests that the clustering pipeline paired with Transformer-based models mostly yield coherent clusters. Coherent, meaning the documents in the clusters are semantically similar as a human interpretable sub-theme of the corpus. E.g. a random sample from a news article corpus will be interpreted as *News*, but the model find coherent clusters of sub-themes such as *Sports*, *Weather*, and events such as phone releases. The Transformer-based language models possess impressive abilities to place semantically similar documents close to each other in a vector space based on the human evaluations in Papers I-III. This means that for a particular document, given enough volume, adjacent documents are likely to be on a similar human-interpreted theme. Using this effect, a clustering algorithm is likely to find coherent clusters by using neighbor-algorithms such as HDBSCAN. In extension, this makes the underlying clusters in BERTopic reliable when used with a Transformer-based vectorization such Sentence-T5 or OpenAI embeddings. The question of whether the topic representations presented to humans (i.e. the keyword representations) are coherent will depend on what algorithm is used for extracting keywords.

The news clusters with lower CIPHE precision are where the clustering algorithm has found a too general cluster such as *Sports*. By collecting articles to generally talk about a subject, humans more often find documents to remove from the cluster. If *Sports* is an adequate granularity level, that should be made clear to the evaluators through acceptance criteria as discussed in Paper V.

In the news domain, most often an event has higher precision according to humans which motivates the use of story-discovery models such as PromptStream (Paper VI, Hatefi et al. (2024)) or micro-clustering (Aggarwal et al., 2003). The main potential improvement area found for the clustering systems worked with here was to disentangle clusters that gradually shifts from one theme to a semantically similar one. E.g. *Movie* clusters often blend with *Celebrities* or *Actors/Actresses* clusters. Since the theme gradually shifts throughout such a cluster, there is no clear separation in the vector space for the dimensionality reduction or clustering algorithms to identify. As discussed by Hatefi (2024), training the vectorization component specifically for the subsequent clustering task through e.g. contrastive learning may be a way forward for specialized tasks such as contextual advertising.

In summary, clustering systems with a Transformer-based language model at their core are likely to produce semantically coherent clusters. These insights are facilitated by CIPHE, which can measure and validate systems for a real-world application such as contextual advertising.

5.4 Future Work

Clustering and topic modeling remain important tools for structuring large corpora of documents when data and information flows are continuously growing. The work in this thesis stands as groundwork for developing a novel evaluation method around the cluster sample validation process, which resulted in the framework CIPHE. CIPHE as a framework has many explorable paths to make it more robust. Mainly, on how information should be conveyed to crowdsource participants. We have discussed three such paths in Section 5.1 regarding the addition of a STELLAR-style corpus overview, the number of documents shown to participants, and the addition of intruder articles. These are all aspects of the framework whose effects should be investigated.

Moreover, the work here was mainly done with crowdsource participants which leaves expert evaluation largely unexplored. Applications of CIPHE to problems inside and beyond the news domain have exciting prospects. CIPHE can be used for validating topic models that are used in digital humanities, validate the contents of clusterings made for a literature summary of a scientific field, or characterizing clusters more accurately in specialized fields such as semantic analysis in reviews or social media. The hope is that applying CIPHE will facilitate deeper understanding of specialized fields and language models.

Chapter 6

Conclusions

This thesis presented methods for evaluating document clusters through human interpretation based on the cluster sample validation process. The development process resulting in the CIPHE framework was discussed, highlighting its capabilities and limitations. The framework was confirmed to function in a crowdsource environment, and to provide characterizations of clusters beyond the capacity of keyword representations. The CIPHE tasks for collecting human interpretation are not reliant on keywords, thus the framework can be used for exploring alternative evaluation strategies for cluster and topic models.

A comparison study of CIPHE and keyword-based methods for collecting human interpretation was carried out in Paper III to investigate the limitations of topic coherence metrics. The study found that keywords was sufficient to represent an overall theme but struggles to represent other characteristics of a cluster. Moreover, the correlation study revealed weak correlations between the collected human interpretation of coherence and the topic coherence metrics. By emphasizing the limitations to keyword-based methods for representing nuances in the underlying documents of clusters and topics, this thesis supports a shift towards more human-centered methods in topic modeling evaluation.

Throughout the thesis, a clustering pipeline consisting of the same building blocks as BERTopic was used for structuring corpora of news articles. Empirical results indicated that the versatile pipeline often yields human interpretable clusters when used with a Transformer-based language model. This was attributed to the high quality of the document embedding vectors, where neighboring documents in a vector space embedded by the same Transformer-model are likely to be semantically similar. For news corpora and contextual advertising, the results indicated that clusters about events had higher CP than clusters with broader themes. This suggests that refining the vectorization process is key to enhancing cluster separability in the vector space in a way that aligns with human interpretation, particularly for domains such as news categorization and contextual advertising.

Finally, this work emphasizes the need for human validation, providing

methods to ensure that systems employing AI, such as Large Language Models, operate in a fair and expected manner. As evaluation complexity is expected to increase with more advanced tasks, human validation remains a crucial endeavor in guiding AI development for the benefit of humanity.

Bibliography

Charu C. Aggarwal, Philip S. Yu, Jiawei Han, and Jianyong Wang. 2003. A framework for clustering evolving data streams. In *Proceedings 2003 VLDB Conference*, pages 81–92. Morgan Kaufmann, San Francisco.

Charu C. Aggarwal and ChengXiang Zhai. 2012. *A Survey of Text Clustering Algorithms*, pages 77–128. Springer US, Boston, MA.

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 13–22, Potsdam, Germany. Association for Computational Linguistics.

Mebarka Allaoui, Mohammed Lamine Kherfi, and Abdelhakim Cheriet. 2020. Considerably improving clustering algorithms using umap dimensionality reduction technique: A comparative study. In *Image and Signal Processing*, pages 317–325. Springer International Publishing, Cham.

Areej Alokaili, Nikolaos Aletras, and Mark Stevenson. 2019. Re-ranking words to improve interpretability of automatically generated topics. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 43–54, Gothenburg, Sweden. Association for Computational Linguistics.

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *Preprint*, arXiv:2008.09470.

R. Bellman, Rand Corporation, and Karreman Mathematics Research Collection. 1957. *Dynamic Programming*. Rand Corporation research study. Princeton University Press.

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2017. An automatic approach for document-level topic model evaluation. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 206–215, Vancouver, Canada. Association for Computational Linguistics.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.

David M. Blei, Michael I. Jordan, Thomas L. Griffiths, and Joshua B. Tenenbaum. 2003a. Hierarchical topic models and the nested chinese restaurant process. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'03, page 17–24, Cambridge, MA, USA. MIT Press.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 113–120, New York, NY, USA. Association for Computing Machinery.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003b. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

T. Caliński and J Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Rob Churchill and Lisa Singh. 2022. The evolution of topic modeling. *ACM Computing Surveys*, 54(10s).

Stephan A. Curiskis, Barry Drake, Thomas R. Osborn, and Paul J. Kennedy. 2020. An evaluation of document clustering and topic modelling in two online

social networks: Twitter and reddit. *Information Processing & Management*, 57(2):102034.

David L. Davies and Donald W. Bouldin. 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Caitlin Doogan and Wray Buntine. 2021. Topic model or topic twaddle? re-evaluating semantic interpretability measures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3824–3848, Online. Association for Computational Linguistics.

Anton Eklund and Mona Forsman. 2022. Topic modeling by clustering language model embeddings: Human validation on an industry dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 635–643, Abu Dhabi, UAE. Association for Computational Linguistics.

Anton Eklund, Mona Forsman, and Frank Drewes. 2023. An empirical configuration study of a common document clustering pipeline. *Northern European Journal of Language Technology*, 9.

Anton Eklund, Mona Forsman, and Frank Drewes. 2024. CIPHE: A framework for document cluster interpretation and precision from human exploration. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 536–548, Miami, USA. Association for Computational Linguistics.

Anton Eklund, Mona Forsman, and Frank Drewes. 2025a. Comparing human-perceived cluster characteristics through the lens of CIPHE: Measuring coherence beyond keywords. *Journal of Data Mining & Digital Humanities*, NLP4DH(32).

Anton Eklund, Albin Nordström, Mona Forsman, and Frank Drewes. 2025b. Industry quality control for efficient continuous human validation of deployed text classification systems. *Submitted to: ACL 2025: Industry track*.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. 2nd Intl. Conf. on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Arezoo Hatefi. 2024. *Deep learning for news topic identification in limited supervision and unsupervised settings*. Ph.D. thesis, Umeå University.

Arezoo Hatefi, Anton Eklund, and Mona Forsman. 2024. PromptStream: Self-supervised news story discovery using topic-aware article representations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13222–13232, Torino, Italia. ELRA and ICCL.

Harold Hotelling. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? The incoherence of coherence. In *Advances in Neural Information Processing Systems*, volume 34, pages 2018–2033. Curran Associates, Inc.

Alexander Miserlis Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2022. Are neural topic models broken? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5321–5344, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2:193–218.

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396–403.

Márton Kardos. 2023. topicwizard: Pretty and opinionated topic model visualization in Python.

Svante Körner and Lars Wahlgren. 2015. *Statistisk dataanalys*, volume 5:3, chapter 11. Studentlitteratur Lund.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden. Association for Computational Linguistics.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *Preprint*, arXiv:1405.4053.

Jia Peng Lim and Hady Lauw. 2023. Large-scale correlation analysis of automated metrics for topic models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13874–13898, Toronto, Canada. Association for Computational Linguistics.

Zachary C. Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Stuart P. Lloyd. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

Jeffrey Lund, Piper Armstrong, Wilson Fearn, Stephen Cowley, Courtni Byun, Jordan Boyd-Graber, and Kevin Seppi. 2019. Automatic evaluation of local topic quality. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 788–796, Florence, Italy. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

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

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, page 262–272, USA. Association for Computational Linguistics.

Fred Morstatter and Huan Liu. 2017. In search of coherence and consensus: measuring the interpretability of statistical topics. *J. Mach. Learn. Res.*, 18(1):6177–6208.

David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. 2010. Evaluating topic models for digital libraries. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, page 215–224, New York, NY, USA. Association for Computing Machinery.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland. Association for Computational Linguistics.

Karl Pearson. 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2.

Erick Peirson, Aaron Baker, Ramki Subramanian, Abhishek Singh, and Yogananda Yalugoti. 2016. Tethne.

Robert-George Radu, Iulia-Maria Rădulescu, Ciprian-Octavian Truică, Elena-Simona Apostol, and Mariana Mocanu. 2020. Clustering documents using the document to vector model for dimensionality reduction. In *2020 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR)*, pages 1–6.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Hamed Rahimi, David Mimno, Jacob Hoover, Hubert Naacke, Camelia Constantin, and Bernd Amann. 2024. Contextualized topic coherence metrics. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1760–1773, St. Julian's, Malta. Association for Computational Linguistics.

Vikas Raunak, Vivek Gupta, and Florian Metze. 2019. Effective dimensionality reduction for word embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 235–243, Florence, Italy. Association for Computational Linguistics.

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

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

G. Salton and M.J. McGill. 1983. *Introduction to Modern Information Retrieval*. International student edition. McGraw-Hill.

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Hendrik Schuff, Lindsey Vanderlyn, Heike Adel, and Ngoc Thang Vu. 2023. How to do human evaluation: A brief introduction to user studies in NLP. *Natural Language Engineering*, 29(5):1199–1222.

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736, Online. Association for Computational Linguistics.

Dominik Stammbach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash. 2023. Revisiting automated topic model evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9348–9357, Singapore. Association for Computational Linguistics.

Michael Steinbach, Levent Ertöz, and Vipin Kumar. 2004. *The Challenges of Clustering High Dimensional Data*, pages 273–309. Springer Berlin Heidelberg, Berlin, Heidelberg.

Nenad Tomasev, Milos Radovanovic, Dunja Mladenic, and Mirjana Ivanovic. 2014. The role of hubness in clustering high-dimensional data. *IEEE Transactions on Knowledge and Data Engineering*, 26(3):739–751.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(95):2837–2854.

Xuerui Wang and Andrew McCallum. 2006. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 424–433, New York, NY, USA. Association for Computing Machinery.

Matthew J Beal Yee Whye Teh, Michael I Jordan and David M Blei. 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad Reza Namazi Rad. 2022. Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3886–3893, Seattle, United States. Association for Computational Linguistics.

He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. 2021. Topic modelling meets deep neural networks: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4713–4720. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel. 2012. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: An ASA Data Science Journal*, 5(5):363–387.