# Comparing Human-Perceived Cluster Characteristics through the Lens of CIPHE: Measuring Coherence beyond Keywords

**Anton Eklund**[1,3], **Mona Forsman**[2], **Frank Drewes**[1]

[1]Umeå University, Sweden
[2]Swedish University of Agricultural Sciences, Sweden
[3]Aeterna Labs, Sweden

Corresponding author: Anton Eklund , anton.eklund@cs.umu.se

**Abstract**

A frequent problem in document clustering and topic modeling is the lack of ground truth. Models are typically intended to reflect some aspect of how human readers view texts (the general theme, sentiment, emotional response, etc), but it can be difficult to assess whether they actually do. The only real ground truth is human judgement. To enable researchers and practitioners to collect such judgement in a cost-efficient standardized way, we have developed the crowdsourcing solution CIPHE – Cluster Interpretation and Precision from Human Exploration. CIPHE is an adaptable framework which systematically gathers and evaluates data on the human perception of a set of document clusters where participants read sample texts from the cluster.

In this article, we use CIPHE to study the limitations that keyword-based methods pose in topic modeling coherence evaluation. Keyword methods, including word intrusion, are compared with the outcome of the thorougher CIPHE on scoring and characterizing clusters. The results show how the abstraction of keywords skews the cluster interpretation for almost half of the compared instances, meaning that many important cluster characteristics are missed. Further, we present a case study where CIPHE is used to (a) provide insights into the UK news domain and (b) find out how the evaluated clustering model should be tuned to better suit the intended application. The experiments provide evidence that CIPHE characterizes clusters in a predictable manner and has the potential to be a valuable framework for using human evaluation in the pursuit of nuanced research aims.

**Keywords**

document clustering, topic modeling, topic modeling evaluation, news clustering, topic coherence, human evaluation methods, crowdsourced cluster validation, BERTopic, CIPHE

## I INTRODUCTION

Document clustering is used to discover patterns in corpora too large to annotate manually. The goal is often to group and label documents by their theme or characteristics, which can be broad categories like Music or narrower ones such as a specific concert. The research field of topic modeling is dedicated to creating algorithms that structure corpora in such a way that the resulting topics and documents belonging to these topics appear interpretable and coherent to a human [Churchill and Singh, 2022, Zhao et al., 2021, Abdelrazek et al., 2023].

Comparing clusters in terms of coherence and human interpretation poses unique challenges. Humans interpret clusters based on their own knowledge and experience, and practical applications of clustering algorithms rarely fit standardized benchmark datasets or gold labelings. For example, whether a news category such as *Technology* should be limited to news about technological gadgets or also encompass news about tech stocks depends on the intended application. The topic modeling field has provided multiple methods based on keyword sets for collecting perceived coherence from human evaluators [Chang et al., 2009, Mimno et al., 2011, Newman et al., 2010b], here called *keyword methods* (KWM, see Section II). The results have been correlated with algorithmic metrics that automatically estimate the coherence of model output without the need for human evaluation [Lau et al., 2014, Röder et al., 2015]. The time and cost efficiency of the latter makes practitioners gravitate towards relying on these algorithmic metrics instead of performing a human evaluation of their models.

Methods based on keywords have been criticized for not capturing the complexities of human interpretation [Doogan and Buntine, 2021] and for relying on statistical word co-occurrence patterns that may not align with human-perceived coherence [Hoyle et al., 2021], a limitation that becomes particularly problematic for neural topic models [Hoyle et al., 2022]. In this study, we address the need for a flexible human evaluation of cluster coherence that does not rely on keywords and can handle semantic document characteristics beyond the main theme. We demonstrate the method CIPHE (pronounced [saıf]) introduced in Eklund et al. [2024], a framework for collecting human interpretation data, and compare it with KWM to highlight in which situations the reduction of clusters to keywords makes the assessment of cluster characteristics unreliable.

CIPHE – *Cluster Interpretation and Precision from Human Exploration* builds on the qualitative approach of simply extracting a sample of documents from a cluster and inspecting them for whatever aspect that might be of interest. To make this doable in a resource-efficient standardized way, CIPHE provides a survey platform and metrics to assess cluster precision and evaluator agreement, as well as human perception of cluster homogeneity and characteristics. Participants of a CIPHE survey provide an interpretation of each cluster by 1) free-text naming a common theme for the majority of articles, 2) identifying texts of the sample that do not fit into the cluster according to this definition, and 3) answering Likert-scale questions about evaluation-specific characteristics. In doing so, participants have access to the full text of the documents they inspect, and can thus assess document characteristics beyond those which keyword lists allow them to discover.

We report on two experiments employing a total of 312 crowdsourced participants. In Experiment I (Section V), we compare the evaluations of 21 document clusters by CIPHE and KWM, and we discuss noteworthy differences. The evaluation covers both cluster coherence and assessments of more intricate cluster characteristics such as whether the documents evoke negative emotional responses. This way, we can highlight cases in which the cluster representation by keywords typically provides sufficient information to assess the desired characteristics and cases where it does not. Experiment II (Section VI) is a case study on a full transformer-based clustering of UK news articles to showcase how CIPHE could be applied in practice and what information it provides about the corpus and clustering model. Finally, we use the results of the experiments to determine the correlation between the CIPHE metrics and common automatic coherence metrics in topic modeling (Section VII).

## II BACKGROUND

Clustering and topic models need to produce output that aligns with human perception of coherence. Methods for measuring and estimating the interpretability of topic model output have long been discussed in the topic modeling field [Wallach et al., 2009, Chang et al., 2009, Doogan and Buntine, 2021, Hoyle et al., 2021]. The human assessment of topic coherence is usually measured indirectly using *intrusion* tasks [Chang et al., 2009]. In the word intrusion task, participants are asked to identify an intruder keyword in a set of keywords representing a topic. Intuitively, if the topic is coherent the intruder will be easy to identify, yielding a higher Mean Precision (MP) on the correctness of the participant choice. Alternative methods for collecting human judgments include direct ones such as having an expert panel rate topics [Mimno et al., 2011] or using crowdsourced workers [Newman et al., 2010b, Aletras and Stevenson, 2013]. However, since studies have demonstrated a high correlation between human judgment and certain algorithmic metrics [Lau et al., 2014, Röder et al., 2015, Newman et al., 2010a], topic models are rarely evaluated by performing any of the human evaluation tasks [Hoyle et al., 2021]. Instead, evaluation is typically conducted through algorithmic metrics such as Normalized Pointwise Mutual Information (NPMI, [Lau et al., 2014, Bouma, 2009]), UMass, and $C_v$ [Röder et al., 2015], which replace the more expensive human evaluations.

Criticism against automatic topic coherence metrics centers around the limitations that the keyword abstraction implies. Doogan and Buntine [2021] argue that automatic coherence metrics do not capture the nuanced ways a topic can be interpreted by humans. Hoyle et al. [2021] criticize such metrics for being overly reliant on word co-occurrence and emphasize that they are not aligned with the context-dependent way in which humans evaluate topics. Building on this, Hoyle et al. [2022] consider the keyword-based evaluation paradigm to be too unstable for the evaluation of neural topic models. Conversely, a participant study by Lim and Lauw [2023] emphasizes that automatic topic coherence metrics are nevertheless meaningful if the right reference corpus is chosen. Notwithstanding, the limiting factor to the current topic modeling evaluation paradigm is that coherence is defined based on keywords.

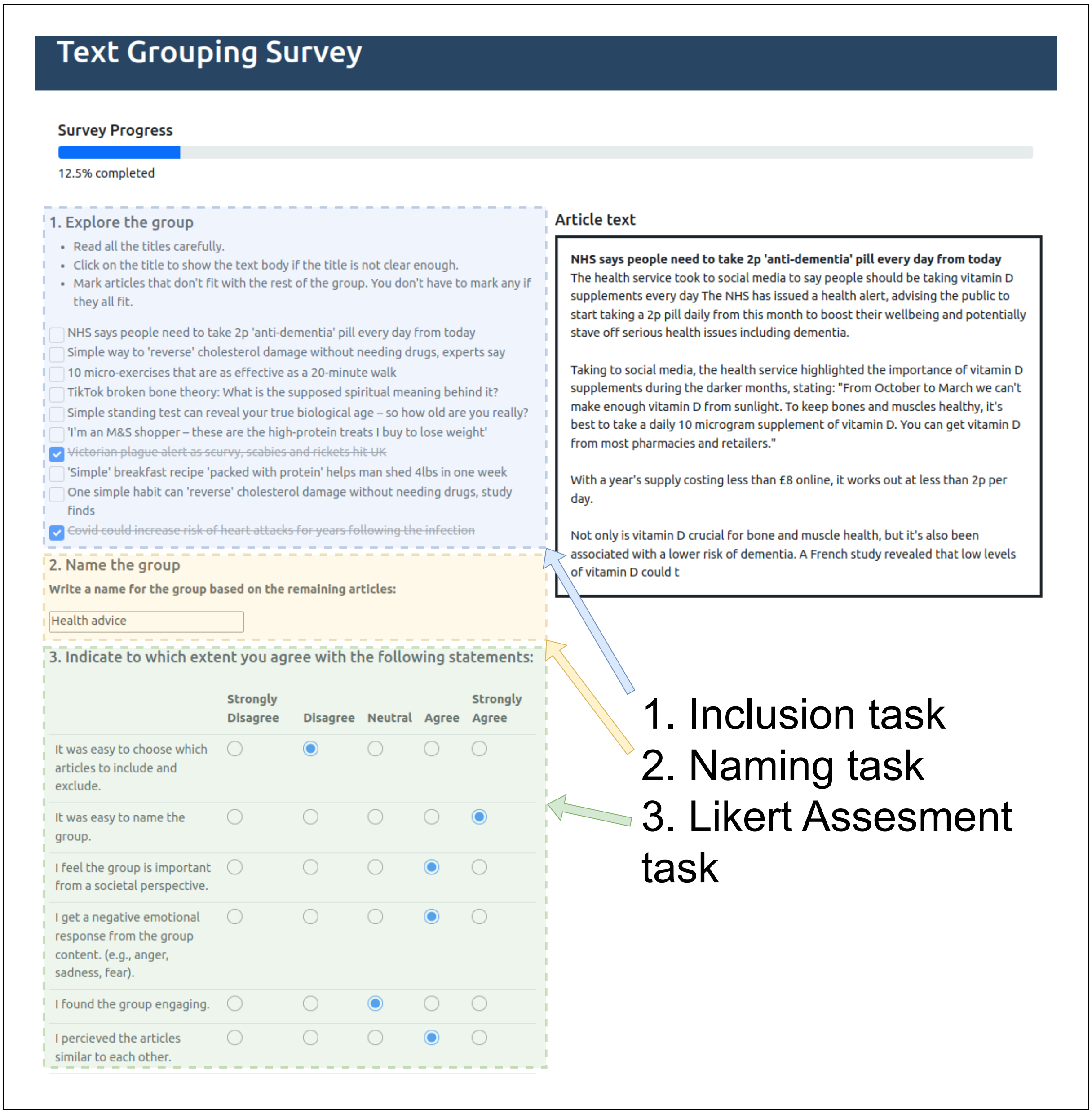
Work has been conducted with the aim to use LLMs to replace the crowdsource workers for the tasks of collecting human interpretation data [Stammbach et al., 2023, Rahimi et al., 2024]. This is a promising avenue forward if it can be confirmed that LLMs capture the document characteristics reflected in human interpretation. To do that, there is a need for collecting human interpretation data beyond keywords, which CIPHE can enable.

In our experiments, KWM use the word intrusion task and the Mean Precision metric (abbreviated MP) by Chang et al. [2009], but also elements similar to the evaluation used by e.g. Newman et al. [2010b] and Mimno et al. [2011] which ask participants to rank topic coherence on a scale.

## III CIPHE

Cluster Interpretation and Precision from Human Exploration (CIPHE) is a framework for collecting human interpretation data of document clusters. The framework consists of a survey platform for data collection (Figure 1), and functions for compiling the data into comparable metrics for the clusters. CIPHE was first introduced in Eklund et al. [2024] where different instruction sets were compared. In this paper, we take into account some drawbacks of the first version and make improvements to the framework to suit a more general use. The framework presented in this paper will stand as the official version moving forward.

CIPHE is built on the assumption that a human who is given a sample of texts from a cluster can use their knowledge and experience to find patterns and answer questions about the sample characteristics without the need to be given a pre-determined list of options common with



Text Grouping Survey

Survey Progress

12.5% completed

1. Explore the group

- Read all the titles carefully.
- Click on the title to show the text body if the title is not clear enough.
- Mark articles that don't fit with the rest of the group. You don't have to mark any if they all fit.

□ NHS says people need to take 2p 'anti-dementia' pill every day from today
□ Simple way to 'reverse' cholesterol damage without needing drugs, experts say
□ 10 micro-exercises that are as effective as a 20-minute walk
□ TikTok broken bone theory: What is the supposed spiritual meaning behind it?
□ Simple standing test can reveal your true biological age – so how old are you really?
□ 'I'm an M&S shopper – these are the high-protein treats I buy to lose weight'
☑ ~~Victorian plague alert as scurvy, scabies and rickets hit UK~~
□ 'Simple' breakfast recipe 'packed with protein' helps man shed 4lbs in one week
□ One simple habit can 'reverse' cholesterol damage without needing drugs, study finds
☑ ~~Covid could increase risk of heart attacks for years following the infection~~

2. Name the group

Write a name for the group based on the remaining articles:

Health advice

3. Indicate to which extent you agree with the following statements:

| | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
|---|---|---|---|---|---|
| It was easy to choose which articles to include and exclude. | ○ | ● | ○ | ○ | ○ |
| It was easy to name the group. | ○ | ○ | ○ | ○ | ● |
| I feel the group is important from a societal perspective. | ○ | ○ | ○ | ● | ○ |
| I get a negative emotional response from the group content. (e.g., anger, sadness, fear). | ○ | ○ | ○ | ● | ○ |
| I found the group engaging. | ○ | ○ | ● | ○ | ○ |
| I percieved the articles similar to each other. | ○ | ○ | ○ | ● | ○ |

Article text

**NHS says people need to take 2p 'anti-dementia' pill every day from today**
The health service took to social media to say people should be taking vitamin D supplements every day The NHS has issued a health alert, advising the public to start taking a 2p pill daily from this month to boost their wellbeing and potentially stave off serious health issues including dementia.

Taking to social media, the health service highlighted the importance of vitamin D supplements during the darker months, stating: "From October to March we can't make enough vitamin D from sunlight. To keep bones and muscles healthy, it's best to take a daily 10 microgram supplement of vitamin D. You can get vitamin D from most pharmacies and retailers."

With a year's supply costing less than £8 online, it works out at less than 2p per day.

Not only is vitamin D crucial for bone and muscle health, but it's also been associated with a lower risk of dementia. A French study revealed that low levels of vitamin D could t

1. Inclusion task
2. Naming task
3. Likert Assesment task

Figure 1: The survey platform as it presents itself to the participants. It shows the titles of the sample documents (1) and, if the participant clicks on one of the titles, the actual text body (to the right). The participant is asked to mark documents that do not fit in (1), to provide a descriptive name (2), and to answer a number of Likert-scale questions (3) regarding their overall impression.

annotation tasks. Given multiple people evaluating the same sample, CIPHE will: 1) estimate the cluster precision from the average number of documents of the sample that the participants considered to actually belong to the cluster, and 2) calculate the similarity of the interpretations of the sample based on the agreement among the responses on the survey tasks and how difficult the participants perceived the performed tasks to be. The CIPHE metrics, which are defined in Section 3.2, are a mix of direct and indirect measurements of cluster quality.

A CIPHE survey requires a group of at least three participants performing the survey. The participants can be crowdsource workers or experts depending on the type of documents and the characteristics to be studied. When the evaluation aim requires collecting data from a general population, a crowdsourcing environment is an appropriate setting.

### 3.1 CIPHE Survey Platform

The CIPHE survey platform is what a participant taking part in a study is exposed to. A CIPHE survey consists of three tasks with general questions that ask the user to assess the cluster sample as follows.

**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.

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

**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 latter ask about characteristics that the investigator has defined, such as negative emotional response, which makes the framework adaptable to different needs.

The survey platform (Figure 1) was implemented in Django[1] and can be found on GitHub at https://github.com/antoneklund/CIPHE/.

### 3.2 Metrics

The metrics applied to the collected responses yield an overall precision estimation for each cluster, reflect different aspects of the agreement between participants, and provide a complexity estimation of the task for each cluster. The purpose of the metrics is to map responses to overall quality scores. CIPHE focuses exclusively on the intrinsic quality of individual clusters rather than assessing a clustering model as a whole, which enables it to evaluate a single cluster in isolation but also implies certain limitations. The upside is that clusters can be evaluated by independent groups of evaluators, and the results reflect the intrinsic qualities of each cluster rather than depending on the context provided by an entire topic model. This also means that clusters from different models can be compared.

### 3.3 CIPHE Cluster Precision (CP)

The precision of a cluster is calculated using the responses from the inclusion task. It is the fraction of the number of documents in the sample which the participants, on average, considered to actually belong to the cluster (according to their interpretation of the cluster).

For each participant $i \in \{1, \ldots, n\}$, let $I_i$ be the set of positive sample documents in $C$, i.e. documents in the sample which participant $i$ considered to belong to cluster $C$. With $m$ denoting the sample size, the *Cluster Precision* of $C$ is

$$\mathrm{CP}_C = \frac{\sum_{i=1}^{n} |I_i|}{nm}.$$

Worth mentioning here is that we have no way of determining the false negatives and calculating the recall, which limits the possibilities of calculating the accuracy of the cluster. This is a consequence of the previously mentioned design decision to evaluate clusters in isolation.

[1] https://www.djangoproject.com/

### 3.4 CIPHE Interpretation and Agreement (IA)

The CIPHE Interpretation and Agreement (IA) is a set of four separate metrics: $\mathrm{A}^{\text{inc}}$, $\mathrm{A}^{\text{name}}$, $\mathrm{L}^{\text{inc}}$, and $\mathrm{L}^{\text{name}}$. IA is focused on determining characteristics of human perceived coherence of clusters and gives a complementary view to CP.

*Remark.* In the previous version of CIPHE [Eklund et al., 2024], IA was a single quantitative metric obtained by taking the average of $\mathrm{A}^{\text{inc}}$, $\mathrm{A}^{\text{name}}$, $\mathrm{L}^{\text{inc}}$, and $\mathrm{L}^{\text{name}}$. This metric was discarded as it seems counter-intuitive to create a summarizing metric for somewhat independent aspects of human interpretation when it cannot generally be determined what is a “good” score for this metric. In particular, a higher score does not always indicate larger coherence.

#### *3.4.1 Agreement Measures*

CIPHE computes two measures of agreement, one each on the inclusion and the naming task.

**Inclusion Agreement $\mathrm{A}^{\text{inc}}$.** The *Inclusion Agreement* metric measures the average pairwise agreement between participants in the decision to exclude individual documents from a given cluster $C$. Equivalently, it can be defined as the average assessment of documents $d$ of the evaluated sample. Thus, in contrast to $\mathrm{CP}_C$, what is measured is not how many documents in $C$ actually belong there, but how well the participants agree on whether or not the documents do.

For the precise definition, recall that the number of (unordered) pairs of elements from a set of size $k$ is $\binom{k}{2} = \frac{k\cdot(k-1)}{2}$. Now, consider first a single document $d$ and assume that its evaluation by participants $1, \ldots, n$ has resulted in $P$ positive votes (and thus $n-P$ negative ones). This means that there are $a = \binom{P}{2} + \binom{n-P}{2}$ pairs of participants who agree in their assessment. The largest possible value is achieved if all $n$ participants have included or all of them have excluded $d$, in which case the result is $a_{\max} = \binom{n}{2}$ whereas the smallest possible value is $a_{\min} = \binom{\lfloor \frac{n}{2} \rfloor}{2} + \binom{\lceil \frac{n}{2} \rceil}{2}$, representing a tie. We thus define the numerical assessment of document $d$ to be

$$\mathrm{A}_d^{\text{inc}} = (a - a_{\min})/(a_{\max} - a_{\min}).$$

This yields $\mathrm{A}_d^{\text{inc}} = 0$ in case of maximum possible disagreement among the participants and $\mathrm{A}_d^{\text{inc}} = 1$ in case of perfect agreement. Finally, the inclusion agreement of $C$ computed on the basis of a sample of $m$ evaluated documents $d_1, \ldots, d_m \in C$ is the average of the agreements on the individual documents:

$$\mathrm{A}_C^{\text{inc}} = \frac{\sum_{i=1}^{m} a(d_i)}{m}.$$

**Naming Agreement $\mathrm{A}^{\text{name}}$.** The *Naming Agreement* reflects the agreement in the free text naming task. To calculate the average agreement on the naming task, we embed the responses with a Sentence-T5-base[2] embedding and calculate the distance between the resulting vectors. This way we measure the semantic similarity of responses rather than their exact formulation. In the case study below, cosine similarity was used as the distance metric. Let $v_1, \ldots, v_n$ be the embedding vectors of the responses of the $n$ participants for cluster $C$ in the naming task and let

$$D_{ij} = \cos(v_i, v_j) = \frac{v_i \cdot v_j}{\|v_i\| \cdot \|v_j\|}$$

for all $i, j \in \{1, \ldots, n\}$. Then

$$\mathrm{A}_C^{\text{name}} = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} D_{ij}.$$

[2] https://huggingface.co/sentence-transformers/sentence-t5-base

#### *3.4.2 Likert Assessments*

The participants evaluating a cluster $C$ are asked to indicate on a Likert scale how much they agree with different statements. Two statements regarding the complexity of performing the inclusion and naming tasks are included in the CIPHE Interpretation and Agreement score of $C$ (Section 3.3). The statements are **inclusion simplicity** ("It was easy to choose which documents to include and exclude") and **naming simplicity** ("It was easy to name the group"[3]). Additional Likert statements can be added to collect data on other cluster characteristics such as emotion, opinion, or bias. The added characteristics and their statements used in this study are described further in Section V.

The Likert scale used for these estimations is $\{$*Strongly Disagree*, *Disagree*, *Neutral*, *Agree*, *Strongly Agree*$\}$. For use in calculations, these responses are converted to the respective numerical scores $0, 0.25, 0.5, 0.75, 1$, and the average over all participants who have evaluated a cluster $C$ is taken. The resulting score is denoted by $\mathrm{L}_C^{\mathrm{inc}}$ and $\mathrm{L}_C^{\mathrm{name}}$, respectively (and similarly for Likert assessments of other characteristics).

#### *3.4.3 Evaluation Sets*

A difference between CIPHE as introduced in Eklund et al. [2024] and the version of CIPHE described here is that it now uses *evaluation sets* to add statistical robustness to the model performance assessment. For this, we use the principle of *Sampling from a Finite Population*, which provides a formula for the minimal sample size $s$ required to determine the proportion of a population which exhibits a certain attribute, depending on the statistical requirements of the investigator [Körner and Wahlgren, 2015].

Let $N$ be the number of documents of a cluster $C$. The required sample size $s$ is then calculated as

$$s = \frac{N\sigma^2}{(N-1)V_0 + \sigma^2} = \frac{N\pi(1-\pi)}{(N-1)V_0 + \pi(1-\pi)},$$

where $V_0$ is the maximum allowed variance of the expected value $v$, and $\sigma^2$ is the population variance in $C$. Since the actual value of $\sigma$ is unknown, it is replaced by $\sigma^2 = \pi(1-\pi)$ where $\pi$ is the estimated proportion of $C$ being correctly classified. The largest $s$ is required when $\pi = 0.5$, which is what we recommend for CIPHE studies. $V_0$ is calculated as

$$V_0 = \left(\frac{\epsilon}{z}\right)^2$$

where $\epsilon$ is the margin of error in the estimation, meaning that the actual value $v$ is in the range $v \pm \epsilon$. For the experiments in this study, we have set $\epsilon = 0.1$ because we are only interested in general tendencies. The confidence level is given by the z-score[4]; throughout this paper we use a $95\%$ confidence level yielding $z \approx 1.96$.

Note that if $N \to \infty$, the sample size will converge to

$$s = \frac{z^2\pi(1-\pi)}{\epsilon^2}.$$

Hence, sample sizes are limited even for very large clusters.

For a CIPHE study, a sample of $s$ documents is randomly selected from each cluster, where $s$ depends on the cluster size as described above. The sample is then partitioned into evaluation

[3] Recall from Figure 1 that the survey uses the non-technical term *group* instead of *cluster*.

[4] see e.g. Mendenhall and Sincich [2007] and, for concrete values, https://www.z-table.com/.

| Dataset | Abbreviation | Documents | Time period | Region | Experiment |
|---|---|---|---|---|---|
| Scraped UK News 1 | SUN1 | 27 786 | 240101–240813 | UK | I |
| Wikipedia Biographies | WIKI | 50 000 | fetched 230424 | Global | I |
| Yelp Reviews | YELP | 50 000 | 050504–220119 | Mainly USA | I |
| Scraped UK News 2 | SUN2 | 22 937 | 241014–241027 | UK | II |

Table 1: Datasets that the clustering model was applied to.

sets of size $m$ each.[5] Each evaluation set is then presented to a number of participants for evaluation. In both experiments of this study, we set $m = 10$ because this number is usually large enough to allow participants to interpret the cluster and is small enough to allow for a quick evaluation and avoid fatigue. Another size of evaluation sets may be chosen if, e.g., the characteristics to be evaluated are very subtle or the evaluation is performed by experts.

To calculate the CIPHE cluster precision CP and the interpretation agreement IA metrics for a cluster, we compute those values for each individual evaluation set from the cluster in question and take their average.

## IV METHOD

This section describes the datasets and methods shared between the experiments.

### 4.1 Datasets

Four different datasets, listed in Table 1, with slightly different characteristics were used in the experiments.

Scraped UK News (SUN1 and SUN2) were included to both try the framework on a real-world use case where standardized taxonomies often are insufficient and to include current affairs which often elicits higher engagement and emotional reactions than old news. SUN1 consists of a mix of daily news and niche news such as tech, food and automotive. SUN2 is limited to a set of the largest news publishers spanning the two weeks prior to the survey to fit the case study in Experiment II where we want to analyze human perception of the current news reporting.

To broaden the perspective of Experiment 1, the datasets WIKI and YELP were added. WIKI contains biographies from Wikipedia [Devinney et al., 2023]. YELP[6] consists of a sample of Yelp reviews in a variety of subjects such as food, veterinarians, or hotels. Common to all the datasets in this paper is that they are considered general enough for crowdsourcing participants to comprehend. For that reason, we do not evaluate on e.g. scientific article abstracts or legal documents in this study.

### 4.2 Cluster Model

For all experiments we used the BERTopic pipeline described by Grootendorst [2022] with the language model Sentence-T5[7] [Ni et al., 2022], UMAP [McInnes et al., 2018] for dimension reduction, and HDBSCAN [Campello et al., 2013] for clustering. The Sentence-T5 model was applied to the dataset without any fine-tuning and the text embeddings were obtained through inference on the title plus text body, without splitting the texts into sentences. The number of

[5] To this end, $s$ is adjusted to the smallest multiple of $m$ at least as large as $s$.
[6] https://www.yelp.com/dataset/
[7] https://huggingface.co/sentence-transformers/sentence-t5-base.

dimensions was reduced to 15 because this has been shown to be a good choice in practice [Eklund et al., 2023]. Further implementation details are found in Appendix A.

The HDBSCAN model produces an outlier cluster of unlabeled documents, which is called *Unlabeled* in all experiments. This cluster consists of articles that are not in sufficient proximity to any of the clusters to be assigned to them according to HDBSCAN. However, it is important to note that these clusters are not really random as they still belong to the overall collection of documents (i.e. news articles, Yelp reviews, and Wiki biographies). They should be seen as sets of documents from those collections taken from less dense regions of the embedding space.

Topic keywords were extracted with cluster-TF-IDF as described by Grootendorst [2022]. It concatenates all documents in a cluster and performs TF-IDF, where the inverse document frequency IDF is determined from the concatenated documents of the other clusters. English stopwords were removed with `nltk`. The top 10 keywords were used to represent clusters to humans in keyword-based surveys and for calculating topic coherence.

### 4.3 Participants

All crowdsource participants were recruited using the Prolific platform[8]. Through the platform, we applied screening for participants to have completed secondary education, and to be born and currently living in the United Kingdom. The reward was set to £10 per hour; the median time for a CIPHE survey was approximately 15 minutes and for the keyword survey around 6 minutes.

## V EXPERIMENT I: COMPARISON WITH KEYWORD-BASED METHODS

Experiment I is designed to make a comparison between CIPHE and KWM to find out in which cases the keyword-based approach suffices and in which it is too limited. As mentioned in Section II, we call the score yielded by the word intrusion task Mean Precision (MP, Chang et al. [2009]). We hypothesize that the reduction of clusters to keywords, while usually appropriate for describing the topic of discussion, is less well suited for identifying characteristics such as sentiment, emotion, and opinion. Experiment I tests this hypothesis by comparing the precision scores and the characterization of clusters of a CIPHE survey with the MP scores and characterization obtained from the corresponding KWM survey.

The characteristics we were interested in for SUN1 were: negative emotional response (Negative Emotion), perceived importance for society (Impact), and engagement in the cluster content (Engagement). These are all important information for decision-makers at companies and in the public sector. Regarding WIKI and YELP, we studied the same characteristics of negative emotion and engagement, but Impact is replaced to better fit the two domains. For WIKI, it is asked if the people belonging to the group are considered to hold leadership positions in society (Societal Leadership). For YELP, it is asked whether the sentiment of positive and negative reviews is considered to be mixed in the cluster (Mixed Sentiment).

To test the appropriateness of KWM in different scenarios, a few clusters from each dataset were chosen whose characteristics keywords, intuitively, should either be more or less well-suited to capture. By contrasting the KWM scores with the CIPHE scores, such hypotheses can then be confirmed or challenged.

[8] Prolific is a platform for recruiting and rewarding crowdsource workers https://www.prolific.com/.

### 5.1 Clusters

The clustering algorithm was applied to SUN1, WIKI, and YELP respectively with the parameters detailed in Appendix A. Six example clusters from each dataset were selected as described above. Table 2 describes the clusters and indicates our reasons for including them, i.e. the characteristics we expected them to have in the eyes of the human evaluators. The unlabeled cluster for each dataset was added to see whether interesting observations would arise from their evaluation.

| **Cluster description** | **Expected characteristics** |
|---|---|
| SUN1: *UK riots* | High engagement and negative emotion |
| SUN1: *Recipes* | High level of coherence |
| SUN1: *American politics* | High impact |
| SUN1: *Phone reviews* | Low engagement |
| SUN1: *Taylor Swift terror threat* | High negative emotion |
| SUN1: *Space* | Low negative emotion |
| SUN1: *Unlabeled* | Mixed news w/o specific thematic focus |
| WIKI: *Religious leaders* | High impact in society |
| WIKI: *Female writers* | Gender aspect implicit |
| WIKI: *Musicians* | High engagement |
| WIKI: *Swimmers* | Very narrow scope |
| WIKI: *Navy officers* | High impact in society |
| WIKI: *Badminton* | Very narrow scope |
| WIKI: *Unlabeled* | Mixed biographies w/o specific thematic focus |
| YELP: *Mexican restaurants* | Low negative emotion |
| YELP: *New Orleans ghost tours* | High engagement |
| YELP: *Veterinarians* | High mixed sentiment |
| YELP: *Reading Terminal Market* | Phenomenon with a proper name composed of ordinary words |
| YELP: *Hotels* | High mixed sentiment |
| YELP: *Negative restaurant reviews* | High negative emotion |
| YELP: *Unlabeled* | Mixed reviews w/o specific thematic focus |

Table 2: Assessment (made by the authors) of cluster characteristics motivating inclusion in Experiment I.

### 5.2 Experimental Setup

The experiment consisted of six surveys, namely one survey per dataset (specified in Table 2) and method (CIPHE and KWM). For each of the resulting surveys, 22 participants were recruited, yielding a total of 132 participants for the experiment. Each CIPHE survey contained one evaluation set per cluster, consisting of 10 randomly sampled articles from that cluster.

Each KWM survey was set up according to the word intrusion task described in Chang et al. [2009], adapted to the CIPHE platform. Instead of viewing the titles and texts, the participants are shown 10 keywords one of which is an intruder randomly selected from the c-TF-IDF keywords of the other clusters. This task differs slightly from the one by Chang et al. [2009] where only 6 keywords were shown. We used 10 instead of 6 keywords to give the participants more context to answer the Likert-scale questions. The keyword sets are shown in a random order to

| **Characteristic** | **Datasets** | **Statement** |
|---|---|---|
| Negative Emotion | All | I get a negative emotional response from the group content (e.g. anger, sadness, fear). |
| Impact | SUN1 | I feel the group is important from a societal perspective. |
| Societal Leadership | WIKI | The individuals belonging to this group are likely to hold leadership positions in society. |
| Mixed Sentiment | YELP | I feel the group contains a mix of both positive and negative sentiment. |
| Engagement | All | I found the group engaging. [Parenthesis added for KWM: "(I want to read the articles/reviews making up this group.)"] |

Table 3: The statements the participants were asked to answer in the Likert-scale questions.

the participants. The intruder keyword is replaced for every participant to avoid the possibility that a single badly chosen intruder word would determine the intrusion score for a cluster.

The Likert statements that the participants needed to consider are shown in Table 3. The questions are the same for CIPHE and KWM except that, for KWM, a clarifying parenthesis was added to the engagement statement since the participant does not have access to the text body.

### 5.3 Results of Experiment I

This section consists of two parts. First, we present and evaluate the results from the scoring metrics from the inclusion task in CIPHE and the intrusion task in KWM. Then, we compare CIPHE and KWM in terms of the Likert-scale questions assessing the cluster characteristics.

#### *5.3.1 Scoring Metrics*

Metrics such as Cluster Precision CP (CIPHE) and Mean Precision MP (KWM), are used for model comparison and algorithmic development of the clustering system. The scores from the two metrics are compared in Figure 2. One can see that CP has a generally higher average of $avg(\text{CP}) = 0.85$ compared to MP $avg(\text{MP}) = 0.63$. While a comparison of the absolute precision between the metrics is not meaningful, we consider CP to be more significant as it indicates how many of the documents of a cluster indeed belong there (according to the

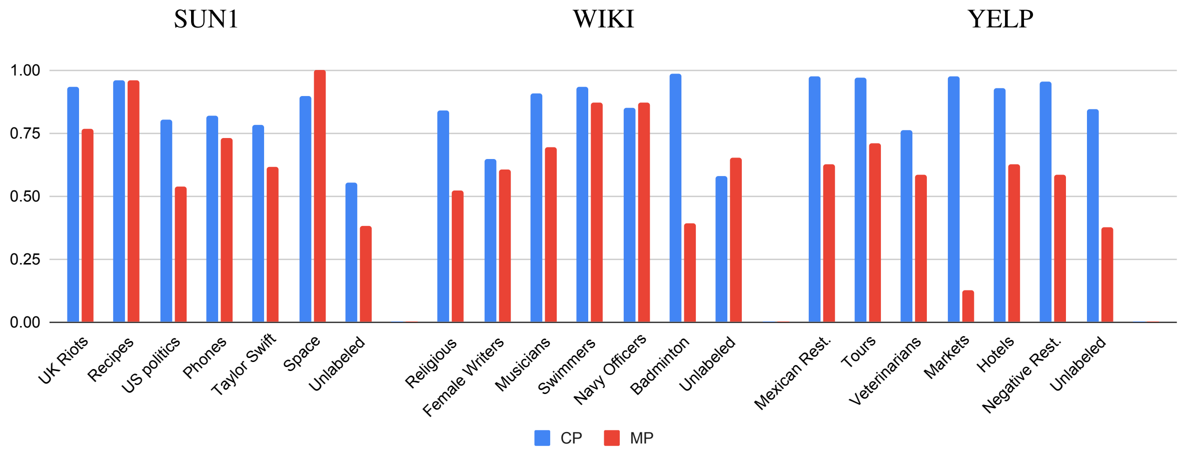


Figure 2: Comparison of CP and MP on the clusters.

| **CIPHE naming** | **KWM naming** |
|---|---|
| Authors/writers (× 6) | American education |
| Authors and publishers | American publishing |
| Campaigners | Award winning children's books by American women |
| Female authors/writers (× 3) | |
| Female activists and politicians | Book publication terms |
| Female novelists/poets | Books (× 2) |
| Females in litersture | Education (× 8) |
| Influential women | Learning |
| Literary figures | Literary |
| Notable figures | Literature (× 2) |
| Philanthropists | Reading |
| Poets/Authors | Study |
| Published authors | Writers (× 2) |
| Women in literature | TBR |
| Writing | |
| Writers and poets | |

Table 4: The free-text naming responses to the cluster *Female writers* by CIPHE and KWM survey participants.

judgement of the participants). Conversely, the MP indicates how often a keyword is missed, which holds less value in the real world.

The *Unlabeled* clusters for all datasets were expected to have a lower precision according to both metrics since they consist of seemingly random documents from each of the datasets. We see that the *Unlabeled* cluster has the lowest CP score for SUN1 and WIKI, but not for YELP. For MP, only the SUN1 *Unlabeled* cluster was ranked lowest. We discuss how the dataset features, specifically that YELP consists of mainly restaurant reviews, interact with the evaluation methods further in Section 8.3.

Pairwise comparison on individual clusters reveals some notable differences in the two scores for some of the clusters. The cluster *Reading Terminal Market* received the lowest MP while CP indicates that it is a highly coherent cluster. The keywords representing the cluster are ⟨`produce`, `like`, `philly`, `vendors`, `amish`, `reading`, `place`, `food`, `terminal`, `market`⟩. We see here that TF-IDF happened to decompose the proper noun "Reading Terminal Market" (in Philadelphia) into a rather incoherent set of independent keywords indicating a bad topic. However, CIPHE participants had a rather easy time realizing that the cluster was indeed coherent. The cluster with a perfect MP score was *Space*, which had the keywords ⟨`spacecraft`, `spacex`, `astronauts`, `earth`, `meteor`, `iss`, `meteors`, `starliner`, `space`, `nasa`⟩. Here, it is rather obvious that almost any intruder would be easy to identify. A third example is provided by the cluster *Female writers* with the keywords ⟨`college`, `family`, `books`, `work`, `award`, `children`, `published`, `american`, `book`, `women`⟩. Here, CP and MP result in similar assessments of cluster coherence. However, looking at the actual cluster names provided by the participants through the respective naming tasks reveals that CIPHE participants understood the cluster to be about *female* writers whereas only one KWM participant noted that fact. Instead, the KWM participants gravitated toward namings about "writers", "education" or "publishing" quite directly derived from one or more of the keywords; see Table 4. These examples illustrate how fragile evaluation with keywords can be and prompt for more robust evaluation frameworks such as CIPHE.

| Dataset | Characteristic | Reject $H_0$ | Fail to reject $H_0$ |
|---|---|---|---|
| SUN1 | All | 8 | 13 |
| WIKI | All | 9 | 12 |
| YELP | All | 14 | 7 |
| All | Negative Emotion | 13 | 8 |
| All | Impact, Societal Leadership, Mixed Sentiment | 10 | 11 |
| All | Engagement | 8 | 13 |
| All | All | 31 | 32 |

Table 5: Results of hypothesis testing using Mann-Whitney U-test. The tests are done per cluster and characteristic, and are aggregated to datasets for readability. The seven clusters of each dataset, with three different characteristics, resulted in 63 comparisons. Rejection of $H_0$ means there is a statistically significant difference between the outcome of CIPHE and KWM. As an example, for the dataset SUN1, $H_0$ was rejected in 8 cases which means that there was a statistically significant difference between the CIPHE and KWM in 8 cases.

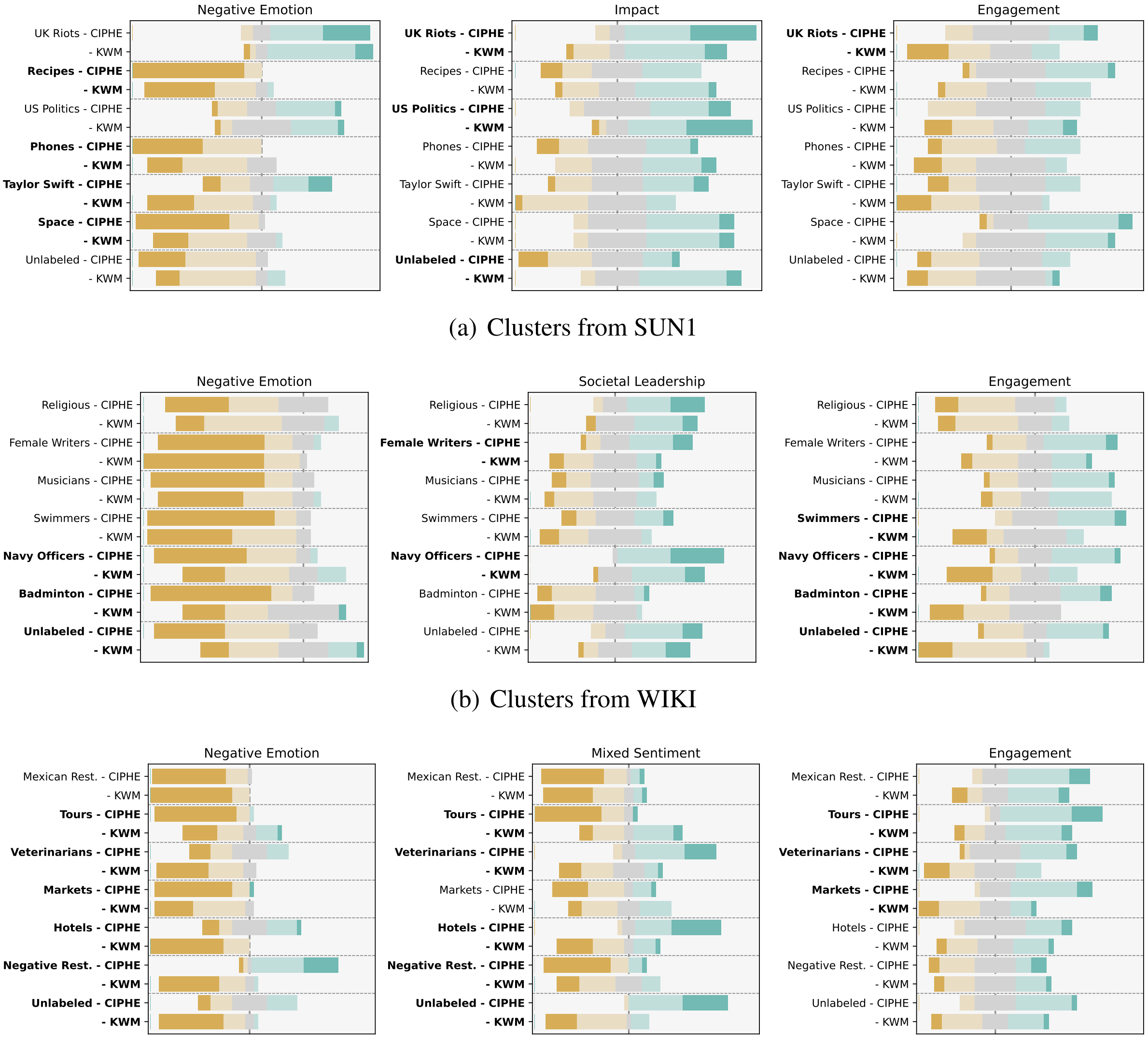
### *5.3.2 Cluster Characteristics*

CIPHE was developed to enable investigators to find deeper characteristics of clusters such as Engagement, Impact, or Negative Emotion. It should be intuitive that abstracting text clusters to keywords results in a loss of the fine-grained information that the texts contain. In Figure 3 the results of the Likert-scale questions are shown. The figures are grouped by characteristic and the outcomes from CIPHE and KWM are paired for each cluster. Additionally, for each cluster and characteristic, the Mann-Whitney U-test was applied to the CIPHE-KWM pairs, where the null hypothesis $H_0$ states that there is no statistically significant difference between the two sets of answers. In the aggregated results in Table 5, we can see that there were significant differences between almost half of the pairs spanning all datasets and characteristics. For the generally rather opinionated YELP dataset, two thirds of the pairs showed a significant difference. This confirms that the keyword abstraction can indeed remove too much information for the participants to be able to properly identify the characteristics of a cluster (depending on the nature of the characteristics considered and that of the clusters).

Let us have a look at some examples of where a high discrepancy between CIPHE and KWM was observed. As expected, the greatest differences occur when some information is hidden in the texts that cannot readily be inferred from the keywords. A prime example of this is the negative emotion connected to the Taylor Swift cluster (Figure 3(a)). Here, the news were about a recent Vienna concert being canceled due to a terror threat, and the related security concerns regarding the upcoming Wembley concert. The only keyword hinting at this was *Vienna*,[9] which holds no negative connotation unless the participant is very well informed. To anyone else, the keyword list is just going to indicate an ordinary cluster about concerts.

For the YELP dataset clusters and the Negative Emotion characteristic, there was a statistically significant difference for all but one pair. For example, the clusters *Veterinarians*, *Hotels*, and *Negative restaurant reviews* contain many negative reviews. Naturally, this evokes a certain negative emotional response, which the CIPHE participants did identify. They also correctly identified that the sentiment was mixed for *Hotels* and *Veterinarians* while *Negative restaurant reviews* only contains negative reviews and is thus not mixed. The cluster *New Orleans ghost*

[9] Full list: ⟨`eras`, `tickets`, `wembley`, `stadium`, `fans`, `tour`, `concerts`, `swift`, `taylor`, `vienna`⟩.



(a) Clusters from SUN1

(b) Clusters from WIKI

(c) Clusters from YELP

Figure 3: Results from the Likert-scale questions, contrasting CIPHE and KWM (shown beneath each other for each cluster). Pairs for which the Mann-Whitney U-test rejected $H_0$ are indicated in bold letters. Scale is ⟨strongly disagree, disagree, neutral, agree, strongly agree⟩ from left to right.

*tours* is the only one for which the KWM results indicated a higher level of negative emotional response than the CIPHE results. The cluster consists of reviews of specifically ghost and cemetery tours in New Orleans, giving rise to keywords such as `ghost` and `haunted`. This exaggerated the negative emotional response in participants seeing only those keywords whereas CIPHE participants considered the cluster to be rather engaging and realized that there were exclusively positive reviews. These examples illustrate that when participants are exposed to the titles and text body, they get a more accurate reading of the cluster content and can make a more reliable assessment than what is possible on the basis of keywords.

The experiment shows that KWM generally works well for finding overall themes, but there is a risk that relevant information is not reflected in the keywords shown to the participant. In particular, KWM struggles with representing semantic aspects beyond those which are implied by the overall theme. CIPHE does not have these problems and allows participants to more accurately characterize clusters with semantic information beyond the main theme.

## VI EXPERIMENT II: CASE STUDY ON FULL NEWS DATASET

CIPHE can be used for evaluating the results of a document clustering and drawing informative conclusions regarding the dataset and the clustering as a whole, as well as individual clusters. To illustrate this, a case study to investigate the UK news domain was prepared. We based the case study on the problem setting of a contextual advertising company that wants to use clustering for keeping track of current events and provide context classification to enable the placement of ads in desired contexts. We consider a news context to be equal to a cluster produced by the clustering model in this study. In this setting, a high CP score indicates that the cluster does not contain unrelated content, which is important to the advertiser. Just as important is the characteristics analysis where information on emotional response, perceived impact on society, and how engaging the cluster content is, could provide useful insights for improving business.

### 6.1 Setup

The dataset SUN2 is a corpus of 22 789 articles from 13 UK publishers collected during the two weeks preceding the participant survey. The clustering algorithm was applied with the settings in Appendix A, resulting in 36 clusters plus the unlabeled cluster of articles deemed outliers by HDBSCAN.

In this experiment we want to evaluate the model accuracy with a higher level of statistical certainty, which means a larger sample of documents per cluster to be evaluated is needed compared to Experiment I. As described in Section 3.4.3, the principle of Sampling from a Finite Population was used to calculate the number of articles needed for each cluster. This resulted in between 50 articles for the smallest clusters and 100 for the largest. The number was rounded up to the nearest factor of 10 for being able to use evaluation sets of size 10. For example, if a cluster had contained 1 000 documents, leading to a required sample size of 88, the sample size would have been rounded up to 90, resulting in 9 evaluation sets with 10 articles each for that cluster.

The evaluation sets were grouped into sub-surveys, each consisting of 8 evaluation sets from different clusters. Each of the 36 sub-surveys was taken by five participants, resulting in a total of 180 recruited participants for the experiment.

### 6.2 Results of Experiment II

The boxplot in Figure 4 shows the median CP together with its standard deviation for each cluster. It is sorted by the median CP with lower values to the right ending with the unlabeled cluster of general world news articles. As a general tendency, on the right side with a median CP of 0.8 one finds broad news categories such as *Weather*, *Food & drink*, and *Music*. Going left towards a median CP of 1 there are more niche news categories and often clusters revolving around particular current events such as the clusters *Death of Liam Payne* or *UK budget submission*. This distribution from left to right suggests that clusters with broader topics tend to have lower precision due to their general nature, while narrower, event-specific clusters achieve higher precision. Therefore, the CP results indicate that, if a higher precision is sought, the broader news categories may need to be sub-divided into smaller ones focusing on current events.

The results of the Likert-scale questions in Table 6 show the highest and lowest average scores for each characteristic. The top clusters for all characteristics were on serious topics such as military conflicts, economy, and health. One outlier to this was that the more leisure-related topic *Food & drink* scored among the highest regarding engagement. The bottom clusters were on entertainment-related topics from areas such as sports, TV, and leisure. From this, we draw

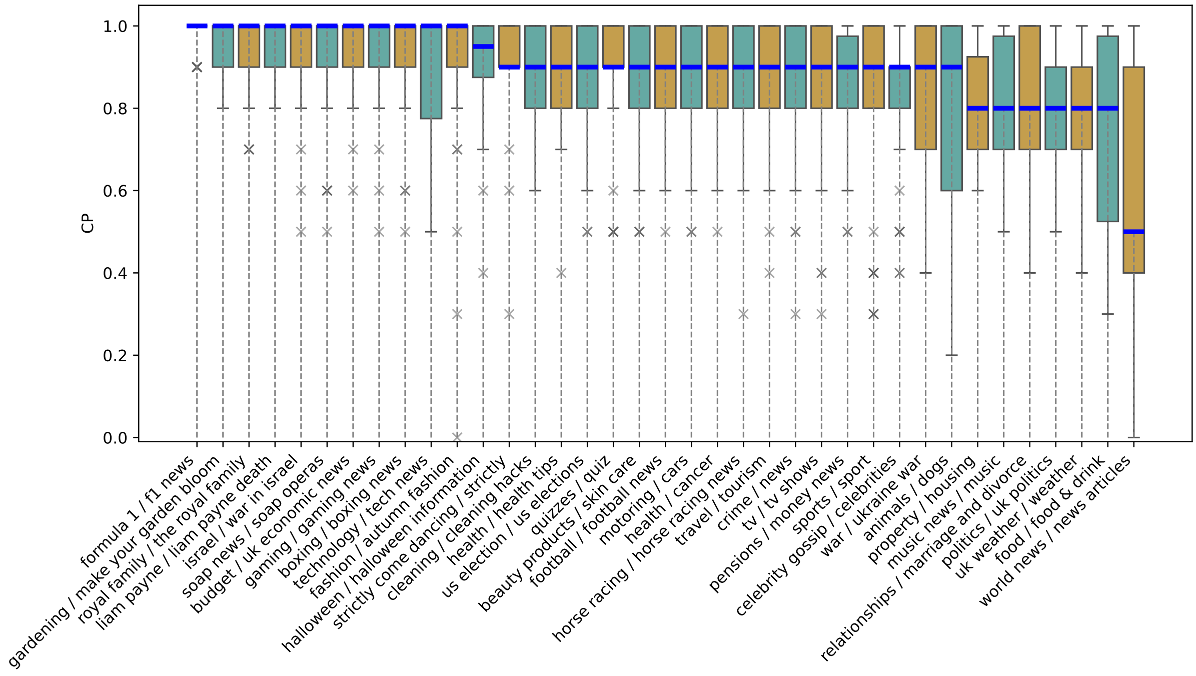


Figure 4: Boxplot for the full evaluation of 37 clusters on a news dataset ranked by their median CP score marked in blue. The cluster names are given from the two most common namings by the participants.

two conclusions confirming what one may intuitively expect. First, more serious topics have a higher potential to cause negative emotion but are also often considered to be more engaging and important for society. From the perspective of effective contextual advertising, it may thus be worthwhile to investigate whether this effect also translates to higher engagement with the advertisement. If so, advertisers should not shy away from these topics. As a side note, it also supports good journalism if publishers can get advertising revenue from serious news. Second, from the perspective of assessing CIPHE as an evaluation tool, it can be said that CIPHE results on cluster characteristics make intuitive sense. Clusters that CIPHE places at the top and bottom with respect to the characteristics in Table 6 are those that should be expected to be found there. At the bottom, we find clusters that are rather niche in their targeted audience compared to news clusters on topics that affect everyone.

## VII CORRELATION ANALYSIS WITH AUTOMATIC METRICS

We finally perform a correlation analysis between the introduced CIPHE metrics with each other and with relevant candidate metrics for automatic topic coherence evaluation. Three common topic coherence metrics ($C_v$, NPMI, and UMASS) were applied to the cluster keywords of the 58 clusters of Experiments I and II. Additionally, the distance-based intrinsic cluster metric *Silhouette coefficient* [Rousseeuw, 1987] was applied to both the original T5 embeddings and the 15-dimensional UMAP-reduced vectors (denoted Sil_768D and Sil_15D, respectively). The topic coherence metrics were calculated using Gensim[10] with 10 keywords per topic, a sliding window of 110 for $C_v$ and UMASS, and with the dataset as the reference corpus. Table 7 shows the correlation matrix obtained by computing the linear correlation Pearson's $r$ between each pair of metrics in the data.

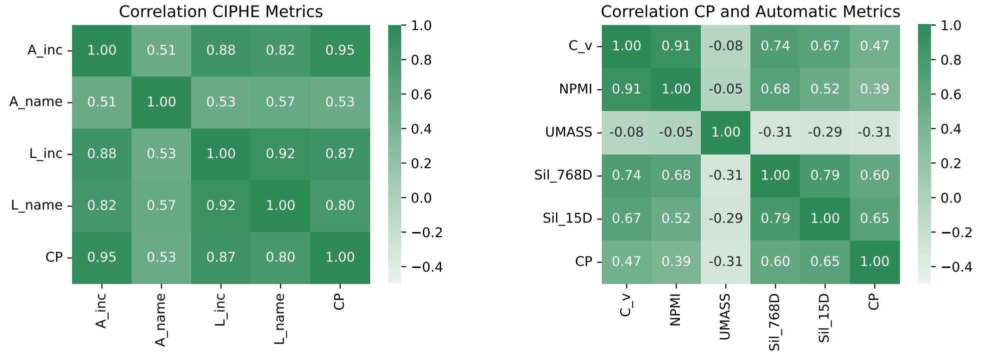
We start by looking at the correlation between the CIPHE metrics CP and the metrics of CIPHE

[10] https://radimrehurek.com/gensim/models/coherencemodel.html

|  | 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 6: Ranked comparison of clusters with respect to Engagement, Impact, and Negative Emotion showing the top and bottom five for each characteristic.

IA, i.e. $A^{\text{inc}}$, $A^{\text{name}}$, $L^{\text{inc}}$, and $L^{\text{name}}$. The correlation matrix (Table 6(a)) shows a high correlation between CP and both of the IA metrics $A^{\text{inc}}$ and $L^{\text{inc}}$. This shows that participants generally find the task more difficult when they need to exclude more articles. Clusters where fewer exclusions are needed naturally also have a higher overall agreement. Among the IA metrics, $A^{\text{name}}$ has the weakest correlation with CP, and is only slightly more correlated with $L^{\text{name}}$. In other words, even if participants largely agree in their view on a cluster, they may come up with semantically rather different names for it (and vice versa). The metric $A^{\text{name}}$ would be interesting to study further, as an extension of it can possibly reveal more precise differences in how participants have interpreted the cluster. Overall, the correlation between CP and IA shows that CP captures many of the other metrics. This is useful for algorithmic improvement as it means that one can focus on reducing the number of documents that participants want to exclude from their evaluation set. However, it is also clear that CP does not capture everything the other metrics measure, meaning that one should be careful making claims on participant agreement or task simplicity based on CP alone.



(a) Intrinsic correlation between the metrics in CIPHE.

(b) Correlation between CP, automatic coherence metrics and the Silhouette coefficients.

Table 7: Correlation matrix for the different metrics, split into two parts. Applied to all evaluated clusters from the datasets SUN1, SUN2, WIKI, and YELP.

Next, we compare CP with the candidates for automatic topic coherence calculation (Table 6(b)). Out of the established topic coherence metrics, $C_v$ has the highest correlation with CP, followed by NPMI, and a negative correlation with UMASS. This is still only a modest correlation and could not be used to reliably estimate the CP score. The correlation between established automatic coherence metrics and MP is lower than reported in previous studies [Lau et al., 2014, Röder et al., 2015]. Using our own results with only the 21 clusters from Experiment I, the correlation of MP with $C_v$, NPMI, and UMASS is 0.45, 0.44, and 0.05, respectively. We attribute this to a more difficult intrusion task (10 instead of 6 keywords) which makes the MP score less stable. Still, our interpretation of these results is that the automatic coherence metrics are indeed measuring the coherence of keyword sets, but should be used with caution when making claims regarding model performance on unseen data.

The two silhouette coefficient metrics have a higher correlation with CP than the coherence metrics. Our previous work on vector embeddings and news article clustering concluded that the embeddings hold the largest influence on the success of a clustering [Eklund et al., 2023]. Similarly, as discussed by Zhang et al. [2022] and many others [Sia et al., 2020, Eklund and Forsman, 2022], a transformer-based language model implicitly structures the vector embeddings of the texts to have close proximity to other similar texts. The high correlation of the silhouette coefficients with CP may indicate that the task of finding coherent clusters could potentially be reduced to finding compact and separated clusters in the vector space. Hence, clustering models may only need to cluster in that vector space to find coherent clusters.

## VIII DISCUSSION

We revisit the aims of the study, which were to identify where KWM falls short, and in the process evaluate CIPHE as a data collection framework for human interpretation of document clusters. This section is structured to start by discussing the findings about how keywords abstract away potentially important information, which has implications regarding the use of automatic coherence metrics in topic modeling evaluation. Finally, we discuss the advantages and limitations of CIPHE as a framework.

### 8.1 Comparison of CIPHE and KWM

CIPHE and KWM behaved notably different with respect to both the resulting precision scores (CP and MP) and how the characteristics described the clusters. Experiment I showed that the variance of CP was lower than that of MP. Generally, low variance in scoring a particular set of example clusters is neither good nor bad, and does not allow us to draw general conclusions. However, a relatively low variance should be expected in Experiment I, because the clusters included in the experiment besides the *Unlabeled* clusters were reasonably well-defined. Hence, the higher variance of MP may be suspected to be an artifact of the particular choice of keywords. One may argue that MP could be made more stable by engineering which keywords can become intruders, determining how many keywords should be shown to the participants, or in other ways making sure that the keyword sets are optimized. Another criticism could be that the c-TF-IDF extraction makes for a low-quality topic model and instead more sophisticated models should be applied to refine the keyword sets. Our standpoint is that future clustering models and topic models, and their evaluation of coherence, should not be based on the optimization of keyword sets. We consider it a strength of CIPHE that it requires less engineering in the evaluation process since the participants are digesting the raw data and making their interpretation based on more information.

Put differently, a CIPHE evaluation of cluster characteristics can be viewed as an approximated ground truth of human interpretation. In fact, if the participants of a survey would have a perfect understanding of the task and they would all evaluate the entire set of documents, the result would by definition be the gold standard of human interpretation. For reasons of practicality, the participants of a CIPHE survey are only exposed to a sample of documents, are not required to read the texts in their entirety, and cannot be assumed to have a perfect understanding of the task. While this means that the results are merely an approximation of the actual ground truth, our experiments show that they are nevertheless trustworthy.

The comparison of the CIPHE and KWM results in Experiment 1 revealed a significant difference in the characterization for 31 out of 63 compared pairs. In the KWM assessment of some clusters, such as *Taylor Swift terror threat* and *Hotels*, the participants missed key characteristics that had been abstracted away by the reduction to keywords. However, KWM were most of the time successful in correctly naming the main theme, meaning that for applications where only the main theme is of interest, keywords still contain enough information. The takeaway is that a small set of keywords has limited capabilities to describe characteristics other than the overall theme.

### 8.2 On the Topic of Coherence

Section II described the background of this work in topic modeling evaluation. What constitutes a topic is vague since it could refer to the word distribution or clusters that a model created, but also hold a definition in everyday speech. If we consider a topic to be a subject of discourse, without taking into account other characteristics such as sentiment, emotion, or style that could be embedded in a document cluster, then the results of this study indicate that keywords are sufficient in most cases for accurately describing the topic of a cluster. If topic modeling is only considered to perform this limited task, then sets of well-chosen keywords and their evaluation may be sufficient. This is where the evaluation of document clusters may differ from topic modeling evaluation. However, we advise against overlooking the potential of using topic or clustering models to examine other, more intricate characteristics.

We saw in the results of the correlation analysis (Section VII) that MP has a weaker correlation with the coherence metrics $C_v$, NPMI, and UMASS than reported in other studies [Röder et al., 2015, Lau et al., 2014]. We attribute this to the larger number of possible choices of keywords (10 instead of 6) in our word-intrusion task. Additionally, the comparatively weak correlations between CP and the coherence metrics indicate that the latter need to be taken with caution. The result from this study support the claims made by Hoyle et al. [2021] and Doogan and Buntine [2021] that the current paradigm of topic coherence metrics based on keywords is not flexible enough to capture the complex nuances of human interpretation. Future work could investigate the correlation between human judgment, such as the CIPHE metrics, and other distance-based automatic candidates such as the silhouette coefficient. This would enable moving away from the dependence on keywords in topic coherence evaluation.

Topic models are applied in many different domains with different requirements and expectations. There will probably never be a single definitive metric that describes how good a clustering or topic model is. That said, pursuing to establish (multiple) reliable automatic metrics for estimating human-perceived coherence is important for efficiency reasons. Recent work on LLM-powered evaluation methods to effectively replace crowdsource workers is an interesting way forward [Rahimi et al., 2024, Stammbach et al., 2023]. CIPHE fits this future as a resource for verifying that LLMs align with human views. Additionally, the CIPHE tasks meet the need for new methods of collecting human interpretation data beyond what the well-established intrusion method can provide [Chang et al., 2009]. The tasks that were performed by crowdsource workers in this study could also be performed by an LLM. However, models ultimately need to be evaluated by humans for making claims about human interpretation or perceived coherence. Such an evaluation could make use of word intrusion in cases where it is applicable, and methods such as CIPHE where characteristics beyond the capabilities of keyword methods are being studied.

### 8.3 CIPHE as a Framework

Experiment II, where CIPHE was used to evaluate the UK news domain and compare the results with those keyword-based word intrusion yields, acts as a demonstration and case study for the framework. Here, we continue the discussion on how CIPHE can be used by practitioners. Experiment II showed that CP declines when the naming of the cluster indicates a broader category. For example, the clusters *Music*, *Weather*, and *Food & drink* all received lower CP scores than more specified ones such as *Formula 1*, *Gardening* and *Royal family*. If a high precision is sought such as in contextual advertisement, the algorithm should be tuned to divide such broad clusters into more specific ones.

Since CIPHE is intended to facilitate collecting interpretation data of characteristics beyond the main theme, there may be numerous adaptions. In Experiment I the Likert questions were adapted to fit specific interests that one might seek from the datasets. E.g. for YELP we asked participants about mixed sentiment and for SUN1 about the importance for society. In such cases, crowdsourcing is likely to be required for recruiting a large pool of evaluators. In other cases, a few domain experts will suffice to perform the evaluation. E.g. evaluating clustering algorithms for scientific literature is not possible to crowdsource to a general audience. Then, a few experts could perform the CIPHE survey but they will each have to evaluate a larger sample. Important here is to use a proper sampling strategy to cover enough documents from each cluster to have reliable outcomes for the research purpose.

From the results of the Likert questions in Experiments I and II we see that CIPHE indeed collects characteristics as intended, where e.g. clusters evoking a negative emotional response are generally those on darker topics. The – in comparison to Eklund et al. [2024] – more carefully formulated Likert-scale questions made it easy for participants to trust their own judgment rather than being reminded of an English exam. Standardizing the inclusion and naming tasks to be the same for any type of data made it easier to formulate survey instructions. The Likert statements are now also easier to adapt to specific research questions as could be seen in Experiment I where one statement was successfully adapted to the different dataset styles.

A limitation of the CIPHE framework was revealed in Experiment I with the YELP cluster. CIPHE identifies the unlabeled clusters of SUN1 and WIKI as low precision clusters, but the unlabeled cluster of YELP was ranked more interpretable than the cluster *Veterinarians*. Upon inspection of the unlabeled YELP cluster, it turned out that it consisted mainly of food reviews as those make up the vast majority of the YELP reviews. In other words, the documents considered to be outliers by HDBSCAN actually constituted a rather coherent cluster as an artifact of the dataset. If a random sample taken from the dataset is likely to show a common theme, CIPHE will most likely not identify it as an algorithmically bad cluster because it actually *is* interpretable to a human. Evidence is provided by looking at the results of the naming task for the unlabeled clusters of the different datasets. The chosen names defaulted to “news articles”, “celebrities”, “restaurant reviews” and the like. These are all derived from properties of the overall dataset as such. If this effect is undesired there is an easy remedy, namely to inform the crowdsource participants about the common characteristics of the dataset and ask them to focus on interpretations of clusters which are less general. Without presenting the participants with this additional information, the observations show that CIPHE works best when the dataset as such cannot easily be mistaken for a meaningful cluster due to some obvious characteristic shared by most documents.

## IX CONCLUSION

We have conducted an in-depth study of the human perception data collection framework CIPHE. CIPHE demonstrated high potential to capture nuanced cluster characteristics, and the flexibility to adapt to diverse research aims. A comparison with keyword-based methods for measuring topic model coherence was made, where CIPHE was able to address certain limitations posed by keywords. The results support criticism towards current standard automatic topic coherence metrics, and we recommend only using them in model development. If making claims on human perceptions of topics (or clusters), it only make sense to validate these claims with human evaluation.

CIPHE stands as an adaptable method for evaluating document clusters, flexible enough to be adjusted to the application environment of the investigator. The case of contextual advertisement studied here as a typical example showcased how the framework can be used to tailor clustering algorithms to the requirements of specific applications. However, there are uncountable other potential use cases for a framework that gathers nuanced characteristics about groups of texts. In a future where model validation is going to be much more demanding because of increased model complexity, human evaluation will become more important than ever.

## DATA AND CODE AVAILABILITY

The code for the CIPHE platform is uploaded at https://github.com/antoneklund/CIPHE/. The articles used in the study and the responses can be provided upon request.

## ETHICS

This study involved the collection of responses through Prolific, a platform where participant identities are known only to Prolific. The survey administered did not include any personal questions and focused solely on annotating the dataset and asking about the complexity of the task. Participants were informed of the purpose of the study and expressed consent for their responses to be used for research purposes. The data collected was securely stored at Umeå University for academic research purposes. Participant anonymity and confidentiality were maintained at all stages of data collection, analysis, and reporting. If participants were to express any concerns or requested their data to be withdrawn, their wishes would be respected without question.

## References

Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. Topic modeling algorithms and applications: A survey. *Information Systems*, 112(C), feb 2023. ISSN 0306-4379. doi: 10.1016/j.is.2022.102131. URL https://doi.org/10.1016/j.is.2022.102131.

Nikolaos Aletras and Mark Stevenson. Evaluating topic coherence using distributional semantics. In Alexander Koller and Katrin Erk, editors, *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 13–22, Potsdam, Germany, March 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-0102/.

Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30: 31–40, 2009.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-37456-2.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei. Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper_files/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf.

Rob Churchill and Lisa Singh. The evolution of topic modeling. *ACM Computing Surveys*, 54(10s), nov 2022. ISSN 0360-0300. doi: 10.1145/3507900. URL https://doi.org/10.1145/3507900.

Hannah Devinney, Anton Eklund, Igor Ryazanov, and Jingwen Cai. Developing a multilingual corpus of Wikipedia biographies. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 285–294, Varna, Bulgaria, September 2023. INCOMA Ltd., Shoumen, Bulgaria. URL https://aclanthology.org/2023.ranlp-1.32/.

Caitlin Doogan and Wray Buntine. Topic model or topic twaddle? re-evaluating semantic interpretability measures. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3824–3848, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.300. URL https://aclanthology.org/2021.naacl-main.300/.

Anton Eklund and Mona Forsman. Topic modeling by clustering language model embeddings: Human validation on an industry dataset. In Yunyao Li and Angeliki Lazaridou, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 635–643, Abu Dhabi, UAE, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-industry.65. URL https://aclanthology.org/2022.emnlp-industry.65/.

Anton Eklund, Mona Forsman, and Frank Drewes. An empirical configuration study of a common document clustering pipeline. In Leon Derczynski, editor, *Northern European Journal of Language Technology, Volume 9*, Linköping, Sweden, 2023. Linköping University Electronic Press. doi: https://doi.org/10.3384/nejlt.2000-1533.2023.4396. URL https://aclanthology.org/2023.nejlt-1.7/.

Anton Eklund, Mona Forsman, and Frank Drewes. CIPHE: A framework for document cluster interpretation and precision from human exploration. In Mika Hämäläinen, Emily Öhman, So Miyagawa, Khalid Alnajjar, and Yuri Bizzoni, editors, *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 536–548, Miami, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.nlp4dh-1.52. URL https://aclanthology.org/2024.nlp4dh-1.52/.

Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*, 2022. doi: 10.48550/ARXIV.2203.05794. URL https://arxiv.org/abs/2203.05794.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. Is automated topic model evaluation broken? The incoherence of coherence. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2018–2033. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/0f83556a305d789b1d71815e8ea4f4b0-Paper.pdf.

Alexander Miserlis Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. Are neural topic models broken? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5321–5344, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.390. URL https://aclanthology.org/2022.findings-emnlp.390/.

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396–403, 2015. doi: 10.9734/BJAST/2015/14975.

Svante Körner and Lars Wahlgren. *Statistisk dataanalys*, volume 5:3, chapter 11. Studentlitteratur Lund, 2015. ISBN 978-91-44-10870-4.

Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In Shuly Wintner, Sharon Goldwater, and Stefan Riezler, editors, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/

E14-1056. URL https://aclanthology.org/E14-1056/.

Jia Peng Lim and Hady Lauw. Large-scale correlation analysis of automated metrics for topic models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13874–13898, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.776. URL https://aclanthology.org/2023.acl-long.776/.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. ISSN 2475-9066. doi: 10.21105/joss.00861.

W. Mendenhall and T. Sincich. *Statistics for Engineering and the Sciences*. Pearson Prentice-Hall, 2007.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In Regina Barzilay and Mark Johnson, editors, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL https://aclanthology.org/D11-1024/.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In Ron Kaplan, Jill Burstein, Mary Harper, and Gerald Penn, editors, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108, Los Angeles, California, June 2010a. Association for Computational Linguistics. URL https://aclanthology.org/N10-1012/.

David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. 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, 2010b. Association for Computing Machinery. ISBN 9781450300858. doi: 10.1145/1816123.1816156. URL https://doi.org/10.1145/1816123.1816156.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.146. URL https://aclanthology.org/2022.findings-acl.146/.

Hamed Rahimi, David Mimno, Jacob Hoover, Hubert Naacke, Camelia Constantin, and Bernd Amann. Contextualized topic coherence metrics. In Yvette Graham and Matthew Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1760–1773, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-eacl.123/.

Michael Röder, Andreas Both, and Alexander Hinneburg. 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, 2015. Association for Computing Machinery. ISBN 9781450333177. doi: 10.1145/2684822.2685324. URL https://doi.org/10.1145/2684822.2685324.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Hendrik Schuff, Lindsey Vanderlyn, Heike Adel, and Ngoc Thang Vu. How to do human evaluation: A brief introduction to user studies in NLP. *Natural Language Engineering*, 29(5):1199–1222, 2023. doi: 10.1017/S1351324922000535.

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.135. URL https://aclanthology.org/2020.emnlp-main.135/.

Dominik Stammbach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash. Revisiting automated topic model evaluation with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9348–9357, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.581. URL https://aclanthology.org/2023.emnlp-main.581/.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 1105–1112, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553515. URL https://doi.org/10.1145/1553374.1553515.

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad Reza Namazi Rad. Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *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, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.285. URL https://aclanthology.org/2022.naacl-main.285/.

He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. Topic modelling meets deep neural networks: A survey. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4713–4720. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/638. URL https://doi.org/10.24963/ijcai.2021/638. Survey Track.

## A IMPLEMENTATION DETAILS

The base Sentence-T5 model from https://huggingface.co/sentence-transformers/sentence-t5-base with simply the function `model.encode(text)` without training or changing any parameters. For SUN1, SUN2, and WIKI, the title and text body of each document were concatenated to a single text that was afterwards, as a whole, encoded by the model. Since YELP does not contain titles only the text body of each document was encoded.

The embeddings were reduced to 15D with UMAP and then clustered with HDBSCAN. The settings for the algorithms depending on the datasets can be found in Table 8. The datasets need different settings to avoid the cluster model resulting in less than 20 clusters.

| Dataset | Dimension | Number of neighbors | Minimal distance | Minimal cluster size | Resulting number of clusters |
|---|---|---|---|---|---|
| SUN1 | 15 | 200 | 0.1 | 60 | 45 |
| WIKI | 15 | 30 | 0.1 | 100 | 58 |
| YELP | 15 | 200 | 0.1 | 100 | 50 |
| SUN2 | 15 | 100 | 0.1 | 100 | 37 |

Table 8: Variable settings for the UMAP and HDBSCAN algorithms.