CIPHE: A Framework for Document Cluster Interpretation and Precision from Human Exploration

Anton Eklund
Department of Computing Science
Umeå University, Sweden
Aeterna Labs, Sweden
antone@cs.umu.se

Mona Forsman
Aeterna Labs, Sweden
mona@aeternalabs.ai

Frank Drewes
Department of Computing Science
Umeå University, Sweden
drewes@cs.umu.se

Abstract

Document clustering models serve unique application purposes, which turns model quality into a property that depends on the needs of the individual investigator. We propose a framework, Cluster Interpretation and Precision from Human Exploration (CIPHE), for collecting and quantifying human interpretations of cluster samples. CIPHE tasks survey participants to explore actual document texts from cluster samples and records their perceptions. It also includes a novel *inclusion task* that is used to calculate the cluster precision in an indirect manner. A case study on news clusters shows that CIPHE reveals which clusters have multiple interpretation angles, aiding the investigator in their exploration.

1 Introduction

Automatically structuring large text collections into clusters is a common research method for its time-saving potential and aiding in discovering patterns. In digital humanities, clustering methods like topic modeling are frequently used for many applications (Newman and Block, 2006; Mimno, 2012; Waheeb et al., 2022; Wallach, 2008; Wickham and Öhman, 2022). Topic models are optimized for structuring texts into coherent themes. However, modern clustering methods powered by sophisticated language models can organize the documents beyond themes. It may be semantic, pragmatic, or other valuable stylistic features. Validating the cluster quality in these cases, or merely discovering such features, requires moving beyond the traditional measures of topic coherence based on keywords (Lau et al., 2014; Röder et al., 2015).

Humans possess a remarkable ability to find patterns, and the discovery of patterns in collections of texts is no exception. Unfortunately, patterns can even be "found" where there are none, a phenomenon called *apophenia* (Shadrova, 2021)[1]. Further, given an overall collection of documents, it is infeasible to objectively define the most appropriate level of granularity in dividing it into topics. Whether *sports* is one news topic or *basketball*, *football*, and *tennis* should be viewed as individual topics depends on the research and application context. Additionally, the background, knowledge, and prior experiences of a reader make it infeasible to establish an objective truth of the cluster properties (Amidei et al., 2019). Thus, researchers and practitioners often focus on specific aspects with carefully formulated questions and have a need to evaluate individual models on unique data.

The connection between topic model output and human interpretation is a topic of debate (Hoyle et al., 2021; Lim and Lauw, 2023; Doogan and Buntine, 2021). Thus, trusting models based on their automatic coherence benchmarking scores may not be good practice for making scientific claims about the data. Moreover, limited quality checks of the topics and apophenia could lead to researchers projecting their own bias to the interpretation of topics, especially if only working with the keyword representation of the topics. To get around this, we suggest performing manual quality validation checks on the actual documents making up a topic. By having human validation of cluster interpretation and precision, there is a stronger basis for making claims based on clustering model results. We propose a framework for collecting data and calculating descriptive metrics for comparing clusters. The framework is aimed toward investigators who either want to systematically validate a model for a specific research question, or who want to use crowdsourcing to collect a general interpretation of a context made up of multiple documents.

A qualitative approach to validating cluster coherence and gaining an understanding of the clusters is to extract a sample of texts from each cluster

[1] *apophenia* - the tendency to perceive a connection or meaningful pattern between unrelated or random things (such as objects or ideas)

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 536–548
November 16, 2024. ©2024 Association for Computational Linguistics

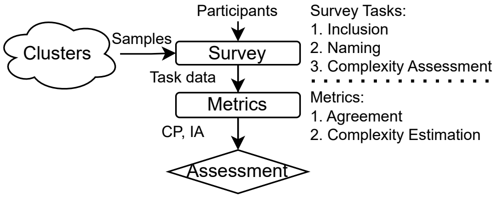


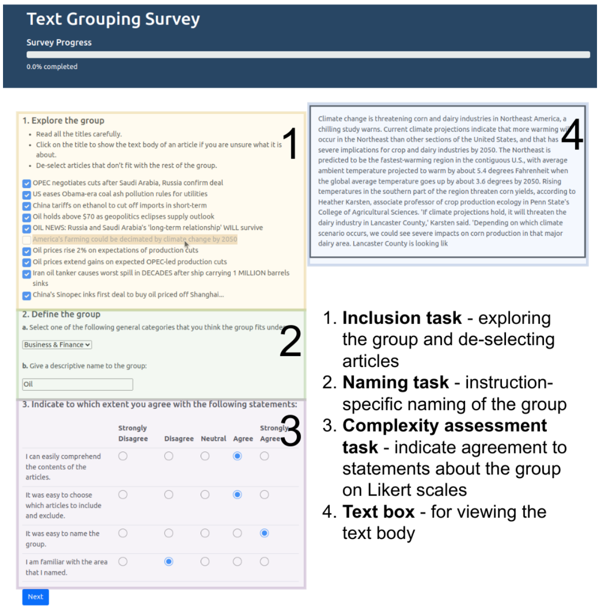
Figure 1: The CIPHE framework.

and inspect them manually, exemplified in Eklund and Forsman (2022). The inspector would then provide an interpretation of each cluster by 1) free-text naming a characteristic feature defining the cluster (i.e., a common theme the majority of its articles belong to), and 2) identifying texts that do not fit into the cluster according to this definition. We propose *Cluster Interpretation and Precision from Human Exploration* (CIPHE) as a framework for recording human interpretation of clusters built around these two tasks. This acknowledges the richness of documents and the possible features to which they can be clustered, and leverages the human ability to recognize patterns to discover cluster properties. This aligns with research requiring interpretative depth and contextual understanding.

This paper consists of two parts. First, we introduce CIPHE as such. Second, we report on a case study with crowdsourcing participants in which CIPHE is examined with respect to its ability to generate valuable insights via crowdsourcing. For this, we conducted a CIPHE survey on ten news article clusters created with different topic models (Section 3). One task of the crowdsource workers is to come up with a descriptive name for each cluster. For this, three sets of instructions were applied and their influence on the task complexity and outcome was discussed. We further analyze the survey results according to the various framework metrics and discuss which insights into the clusters they reveal.

2 CIPHE

To bring structure to the perception of multiple evaluators, we introduce Cluster Interpretation and Precision from Human Exploration (CIPHE, Figure 1) as a framework for recording and analyzing human interpretations of document clusters. Provided that a cluster can be characterized by a feature that most of the texts have in common (which may but does not necessarily have to be a general theme), we expect that a human exploring a sample of the cluster will be able to 1) name this feature, and 2) identify which of the articles do not share it and have thus wrongly been placed into the cluster. The central idea of CIPHE is to exploit these abilities to obtain insights into the quality and properties of text clusters. For this, CIPHE lets a group of survey participants perform both of the mentioned tasks and compares the individual interpretations and quality assessments. A CIPHE survey can be performed by either a small group of experts or a larger number of participants. The large number of participants accessible via crowdsourcing allows human evaluation to be based on a broader interpretation of texts and can mitigate certain biases (Schuff et al., 2023) while an expert survey can be used if the clustered documents or the demands on the clustering require expert knowledge.


Text Grouping Survey

Survey Progress

0.0% completed

1. Explore the group

- Read all the titles carefully.
- Click on the title to show the text body of an article if you are unsure what it is about.
- De-select articles that don't fit with the rest of the group.

☑ OPEC negotiates cuts after Saudi Arabia, Russia confirm deal
☑ US eases Obama-era coal ash pollution rules for utilities
☑ China tariffs on ethanol to cut off imports in short-term
☑ Oil holds above $70 as geopolitics eclipses supply outlook
☑ OIL NEWS: Russia and Saudi Arabia's 'long-term relationship' WILL survive
☐ America's farming could be decimated by climate change by 2050
☑ Oil prices rise 2% on expectations of production cuts
☑ Oil prices extend gains on expected OPEC-led production cuts
☑ Iran oil tanker causes worst spill in DECADES after ship carrying 1 MILLION barrels sinks
☑ China's Sinopec inks first deal to buy oil priced off Shanghai...

2. Define the group

a. Select one of the following general categories that you think the group fits under:

Business & Finance

b. Give a descriptive name to the group:

Oil

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

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I can easily comprehend the contents of the articles.	○	○	○	●	○
It was easy to choose which articles to include and exclude.	○	○	○	●	○
It was easy to name the group.	○	○	○	○	●
I am familiar with the area that I named.	○	●	○	○	○

Next

Climate change is threatening corn and dairy industries in Northeast America, a chilling study warns. Current climate projections indicate that more warming will occur in the Northeast than other sections of the United States, and that has severe implications for crop and dairy industries by 2050. The Northeast is predicted to be the fastest-warming region in the contiguous U.S., with average ambient temperature projected to warm by about 5.4 degrees Fahrenheit when the global average temperature goes up by about 3.6 degrees by 2050. Rising temperatures in the southern part of the region threaten corn yields, according to Heather Karsten, associate professor of crop production ecology in Penn State's College of Agricultural Sciences. 'If climate projections hold, it will threaten the dairy industry in Lancaster County,' Karsten said. 'Depending on which climate scenario occurs, we could see severe impacts on corn production in that major dairy area. Lancaster County is looking lik

1. **Inclusion task** - exploring the group and de-selecting articles
2. **Naming task** - instruction-specific naming of the group
3. **Complexity assessment task** - indicate agreement to statements about the group on Likert scales
4. **Text box** - for viewing the text body

Figure 2: The Survey platform. We use the non-technical term *group* instead of *cluster* for ease of interpretation by the participants.

The quality assessments are indirect, which makes them more comparable: rather than asking a participant explicitly to assess the quality of a cluster on a subjective scale, quality is inferred from the number of documents they exclude from the cluster. In general, different participants may name clusters differently, and their segmentation, i.e. which articles they choose to include in their interpretation of the cluster, will also vary. Rather than viewing this as a problem, CIPHE builds upon it. For a given cluster, the degree of agreement between participants and their individual assessments of the task complexity are converted to comparable metrics for cluster quality (Section 2.1).

A CIPHE survey consists of three tasks that collect responses reflecting the interpretation of the cluster by the participant.

Inclusion: the participant is asked to explore the cluster and decide which articles, according to them, belong to the cluster. Intuitively, the fewer articles are excluded, the better the cluster in the eyes of the participant.

Naming: the participant is asked to give the cluster a descriptive free text name. The precise instructions for how to do this may differ. In our case study, we compare three different instruction sets; see Section 3.3.

Complexity assessment: the participant answers Likert-scale questions about their experience exploring the cluster (Joshi et al., 2015). This provides information about both the participant and the perceived simplicity of interpreting the cluster.

A survey platform (Figure 2) was implemented in Django[2] to be able to manage the survey in detail and to have control over how the different elements were displayed to participants.

2.1 Metrics

The metrics applied to the responses were chosen to 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. The current version of CIPHE focuses exclusively on the intrinsic quality of individual clusters rather than assessing a clustering model as a whole, making it applicable when working with a single dataset and model.

2.1.1 Agreement Measures

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

Inclusion Agreement $\mathrm{A}^{\mathrm{inc}}$: The *Inclusion Agreement* metric measures the pairwise agreement between participants in the decision to include or exclude individual documents to the cluster. This metric is robust to participants having diverging views for a few individual documents in the cluster but mostly agree on the rest.

Let the sample of documents from cluster C be $d_1, \dots, d_m$. For participants i and j, let $\mathscr{A}_{ij}$ be the set of all d_k, $k \in \{1, \dots, m\}$, on which i and j agree, i.e. either both have included d_k in C or both have excluded d_k from it. Then

$$\mathrm{A}_C^{\mathrm{inc}} = \frac{2}{mn(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} |\mathscr{A}_{ij}|$$

where $n > 1$ is the number of participants.

	Part. A	Part. B	Part. C	Part. D
Doc 0	i	i	e	e
Doc 1	i	i	i	i
Doc 2	e	e	e	i
Doc 3	i	i	e	e

Table 1: Example with a set of four documents and four participants with the decision denoted i and e for including or excluding, respectively. Participant A agrees with B on all documents, with C on two documents, with D on one. Participant B agrees with C on two documents and with D on one. Participants C and D agree on 3 documents. This gives an $\mathrm{A}^{\mathrm{inc}}$ score of $\frac{2}{4 \cdot 4 \cdot (4-1)} \cdot (4+2+1+2+1+3) = \frac{13}{24}$. Participants A and B have made identical segmentations of the documents, and C and D have made individual segmentations. The resulting $\mathrm{A}^{\mathrm{seg}}$ score is $1 - \frac{3-1}{4-1} = \frac{1}{3}$.

Segmentation Agreement $\mathrm{A}^{\mathrm{seg}}$: The *Segmentation Agreement* measures the participant agreement on how to segment the documents into two sets: included and excluded documents. A high $\mathrm{A}^{\mathrm{seg}}$ means that participants more frequently have chosen the same set of documents to include, implying that there are few ambiguous documents in the cluster. In contrast to $\mathrm{A}^{\mathrm{inc}}$, the metric is sensitive to small differences in decisions on individual documents, as they create new segmentations. See the example of $\mathrm{A}^{\mathrm{inc}}$ and $\mathrm{A}^{\mathrm{seg}}$ in Table 1.

Again, let $d_1, \dots, d_m$ be the sample of cluster C and assume that there are n participants. For $i \in \{1, \dots, n\}$, let I_i be the set of documents d_i which, according to participant i, indeed belong to C. Let $u = |\{I_i \mid 1 \leq i \leq n\}|$, i.e. u is the number of unique segmentations of C (the number of ways the participants have divided C into). Then

$$\mathrm{A}_C^{\mathrm{seg}} = 1 - \frac{u-1}{n-1},$$

yielding a score of 0 if all participants disagreed and 1 if they all agreed.

Naming Agreement $\mathrm{A}^{\mathrm{name}}$: 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[3] embedding and calculate the

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

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

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

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

The normalization constant λ is experimentally set to 0.6 to increase the variance of A^{name} and thus its impact in subsequent calculations. Experiments showed that $\min(D_{ij})$ was larger than 0.65 after pairwise comparison between all responses under all instructions.

2.1.2 Complexity Estimation

In the complexity assessment task, the participants are asked to indicate on a Likert scale how much they agree with different statements regarding the survey task. This is not primarily to estimate the difficulty of the survey itself, but mostly to gain insights into the perceived simplicity and coherence of clusters.

Each participant is asked to provide an estimate of the level of agreement with statements regarding **comprehension** ("I can easily comprehend the contents of the articles"), **inclusion simplicity** ("It was easy to choose which articles to include and exclude"), **naming simplicity** ("It was easy to name the group")[4], and **knowledge** ("I am familiar with the area that I named").

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 $0, 0.25, 0.5, 0.75, 1$. Neither *comprehension* nor *knowledge* is used in CIPHE quality metrics. They were added because they may provide useful insights for additional targeted evaluations.

Let $\mathscr{L}_i^{\text{inc}}$ and $\mathscr{L}_i^{\text{name}}$ be the numerical values of the responses of participant i ($i \in \{1, \ldots, n\}$) to the Likert inclusion and naming statements, respectively, for cluster C. Then the metrics $\mathrm{L}_C^{\text{inc}}$ and $\mathrm{L}_C^{\text{name}}$ are calculated as

$$\mathrm{L}_C^{\text{inc}} = \frac{1}{n} \sum_{i=1}^{n} \mathscr{L}_i^{\text{inc}} \quad \text{and} \quad \mathrm{L}_C^{\text{name}} = \frac{1}{n} \sum_{i=1}^{n} \mathscr{L}_i^{\text{name}}.$$

2.1.3 Cluster Quality Metrics

CIPHE Precision CP: The precision of a cluster is calculated using the responses from the inclusion task. For each participant $i \in \{1, \ldots, n\}$, let I_i again 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 *CIPHE 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.

CHIPE Interpretation and Agreement IA: The CIPHE interpretation agreement score is the average of the sum of all agreement and complexity estimation scores:

$$\mathrm{IA}_C = \frac{\sum_{a \in A} \mathrm{A}^a + \sum_{\ell \in L} \mathrm{L}^\ell}{|A| + |L|}$$

where $A = \{inc, seg, name\}$, $L = \{inc, name\}$.

3 Case Study

A case study was conducted to validate the ability of CIPHE to quantitatively differentiate clusters in terms of interpretation and precision using human interpretation data collected via crowdsourcing.

3.1 Dataset

We selected clusters of varying quality to evaluate CIPHE in different situations. Four clusters were selected from the well-curated WCEP18 corpus (Yoon et al., 2023), and six were from a less polished scraped dataset of English news articles from 2022 that gives a more realistic view of a real-world application.

3.2 Clusters

The topic models Latent Dirichlet Allocation (LDA Blei et al. (2003)), BERTopic (Grootendorst, 2022), and the story discovery model PromptStream (Hatefi et al., 2024) were applied to WCEP18 resulting in 150 topics for LDA, 318 topics for BERTopic

[4] Recall from Figure 2 that the survey uses the non-technical term *group* instead of *cluster*.

No.	Name	Characteristics	Expected quality	Reason for inclusion
1	NFL BLM Protest	The event where NFL players took a knee for Black Lives Matter and the political aftermath.	High	Random cluster from models
2	South Africa	South Africa Politics and international news about land rights for farmers.	High	Random cluster from models
3	Financial Advice	Different articles on how to save money for individuals. It contains many different angles on this subject such as mortgage, collecting Covid support checks, pension, and credit card comparisons.	Medium	Diverging internal views
4	Macedonia Name Protest	Greek Protest about Macedonia changing their name. Also includes two irrelevant articles about a hostage named Joshua Boyle.	High	Random cluster from models
5	Oil	News about oil prices but also contains energy and environment.	High	Random cluster from models
6	Celebrities	The cluster contains articles that could be divided into many different segments depending on the knowledge of the participant. Gossip, celebrities, family, Reality TV, entertainment.	Medium	Diverging internal views
7	Tips and Tricks	A cluster that follows a pattern of the title containing "I am a . . . " and then proceeds to advise on a wide range of fields. E.g. "I'm an interior design expert – 3 easy ways to make your home look way more expensive on the cheap."	Medium	Diverging internal views
8	Astronomy	Articles about space and meteors. But also contains 3 articles about animals and bugs.	Medium	Diverging internal views
9	Cannabis/IT Security	Artificially created cluster by combining five articles from two distinct clusters which were Cannabis legalization and IT security leaks.	Low	Test participant reaction to clusters that combine distinct topics
10	Random	A cluster of random articles. The model grouped them due to similar article lengths.	Low	Baseline & estimate apophenia effects

Table 2: The clusters used in the case study, ranked by Topic Coherence metric c_v (Röder et al., 2015). Expected quality was estimated by the authors prior to releasing the survey.

and 525 stories for PromptStream. Four clusters were randomly chosen from the resulting pool of model outputs. We (the authors) determined these to be of high quality so to include clusters of varying quality and corner cases, six clusters from the scraped dataset were added. Four clusters where we had diverging views on how to characterize the cluster, despite agreeing that the cluster was reasonably well defined. One cluster comprised of randomly chosen articles to have a baseline and to be able to estimate the influence of apophenia. Lastly, one artificially constructed cluster by combining equal numbers of articles belonging to two distinct topics, to be able to evaluate the answering patterns of participants in this artificial corner case.

A sample of ten articles was extracted from each cluster. The same ten articles are shown to every participant.[5] The detailed cluster descriptions can be seen in Table 2.

3.3 Instruction Sets

Three different instruction sets were used for the naming task which vary in their degree of freedom to interpret the cluster. These were:

Free-text (FT): The participants were simply asked to name the cluster. This gives the largest degree of freedom. The expected outcome from using this instruction set was to get specific names, but also with semantic diversity due to the creativity and different perspectives of the participants.

Unifying Features (UF): The participants were instructed to first choose whether the cluster was about an *event*, *general theme*, or *something else*, and then specify in free-text. The minimal initial structuring that this provides was meant to prime the participant for more descriptive naming. This aids with determining which articles should not have been included in the cluster and understanding clusters that may seem incoherent initially. UF provides a large degree of free human interpretation beyond a pre-defined taxonomy, but is limiting by making the participants precede their decision by a

[5] Note that the sample size of 10 was chosen to evaluate CIPHE in a controlled setting. For an actual evaluation, multiple samples of articles from the same cluster are required to reliably characterize the cluster.

high-level classification.

Taxonomy (TAX): The participants are given a taxonomy to choose an overall news category (similar to annotating a dataset for classification) and are then asked to name the cluster in free text. This is a low degree of freedom in the first step, but anticipated to add specificity in the second.

The focus of UF on themes and events was chosen because this case study works with news articles. The same holds for the taxonomy created for TAX. For other types of data, this may need to be adjusted. In contrast, the instruction set FT is universally applicable.

3.4 Participants

The experiment involved 20 participants for each of the 3 instruction sets, giving a total of $N = 60$ participants. The participants where recruited in Prolific[6] using their standard sample. The only requirements were that the participants should be fluent in English, and have graduated from secondary education. We deliberately did not control for other demographic parameters because we wanted to capture as general a set of views as possible. This also limits the extent to which we can analyze the influence of the background of participants on the responses given. A detailed demography of the participants can be seen in Appendix A. The participants were paid £10/h for approximately 25 minutes of work. Due to some participant responses having too low quality, i.e., the participant did not exclude a single article for the duration of the survey, or otherwise clearly misinterpreted the instructions, we recorded that as an instruction failure, and recruited replacement participants. For each instruction set, 2 instruction failures were recorded.

3.5 Survey

The participants were informed about the general goal of the study and asked for consent (Appendix B.1). Then, they received one of the three sets of instructions (Appendix B.2) and proceeded to the survey question pages. The survey platform (Section 2) displayed one random cluster at a time to participants, starting with a cluster from WCEP18. The decision to always start the evaluation with a cluster from WCEP18 was made after a pilot study showed that participants had difficulties understanding the survey instructions when starting with the *Random* or the artificially created clusters.

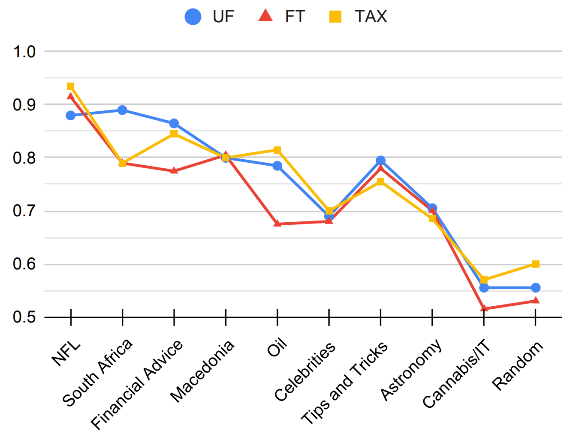


Figure 3: The CP metric for the tree instruction sets.

After the survey was completed we assessed the answers. If there were missing answers or signs of technical problems, the participants were asked to retake parts of the survey.

4 Results and Discussion

CIPHE is evaluated on its ability to capture the varying quality of the clusters and human interpretation of them (Section 4.1). In our analysis, we use the fact that some of the manually selected clusters, the artificial cluster, and the random cluster should be more difficult to interpret. Further, we compare the instructions and analyze their effect on the responses (Section 4.2). We also discuss adaptions that can be made to suit different usage purposes (Section 4.3).

4.1 Metric Analysis

4.1.1 Cluster precision CP

The cluster precision, CP, calculates the average ratio of included articles in a cluster and functions as an indirect measurement for cluster coherence. The scores (Figure 3) range between 0.52 and 0.94 which shows that participants on average view between five to nine articles as correctly belonging to a cluster. Both clusters with expected higher and medium quality were found in the mid-range from 0.68 to 0.85. The cluster *Random* and the artificial cluster *Cannabis/IT* have the lowest scores which shows that CP correctly identifies improvement areas in the clustering model performance. *Random* scored between 0.53 and 0.60 which is higher than anticipated. This may be a combination of apophenia, and that many participants (correctly) interpreted the cluster as general news (see Table 3).

[6] https://www.prolific.com/

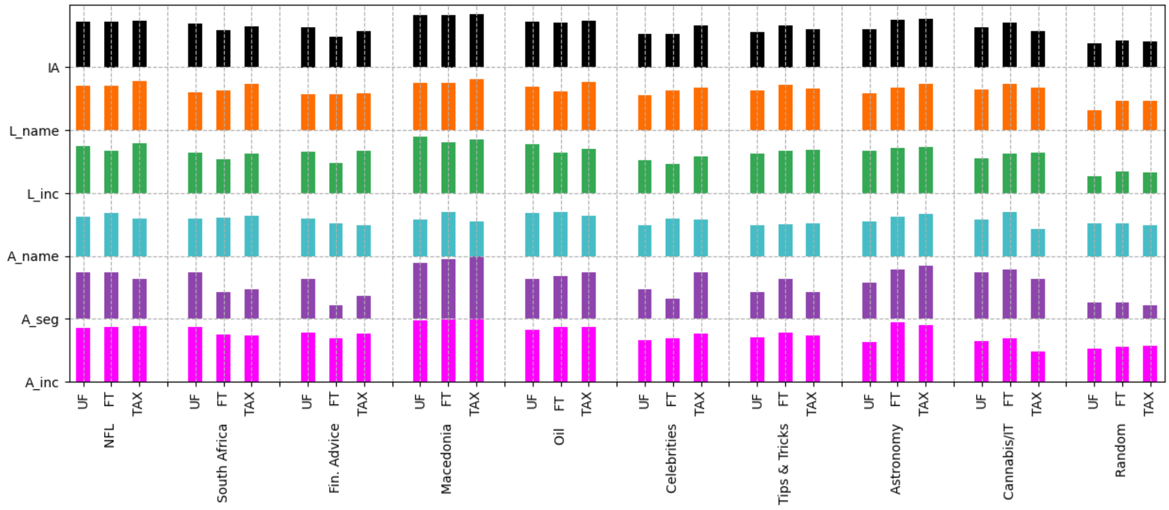


Figure 4: Metrics A^{name}, A^{seg}, A^{inc}, L^{name}, L^{inc}, and IA for each cluster and instruction set. Each metric is bound between $[0, 1]$ and should be compared horizontally. The instructions UF, FT, and TAX can be compared by inspecting differences for a cluster found on the x-axis for each metric.

The results also show that participants feel inclined to exclude at least one article as no cluster received a score of 1.0. A potential measure for mitigating this effect is discussed in Section 4.3.

While CP does not reveal deeper insights about the cluster, it gives each cluster an intuitive quality score connected to human perception usable for algorithmic improvement in many applications. We note that the CP scores were only slightly influenced by the choice of instructions. This indicates that the inclusion task of CIPHE itself is able to capture cluster quality in an indirect manner, making it a promising alternative for e.g., collecting human interpretations as a basis for improving topic models (Chang et al., 2009; Newman et al., 2010).

4.1.2 Agreement Measures

The metrics on data from the inclusion task of Inclusion Agreement (A^{inc}) and Segmentation Agreement (A^{seg}) explain much of the variance between instructions within the same cluster in Figure 4. A^{inc} measured participant consensus on including individual articles, scoring higher in clusters with obvious outliers like *Macedonia*, but generally correlated with the CP score. A^{seg} revealed when participants had multiple unique segmentations for a cluster, with lower scores indicating potential ambiguity, as seen in e.g., the *Financial Advice*, *Tips & Tricks*, and *Random* clusters. However, A^{seg} can be somewhat volatile as it in many instances has a high variance in scores between the instructions on the same cluster.

The differences between clusters in the naming agreement A^{name} were low due to participants writing free-text answers that group closely together in an embedding space. Overall, the participants named the clusters similarly (Table 3) which indicates that they identified similar broader topics, even if they chose different words to specify them. The FT instruction had slightly higher A^{name} scores which we attribute to it prompting participants to answer in a few words, a prompt that is not part of the TAX or UF instruction sets.

4.1.3 Complexity Estimation

The complexity estimation metrics are collected with Likert-scale questions asking the participants about the simplicity of performing the inclusion and the naming task. The results from L^{inc} and L^{name} in Figure 4 most closely resembled our initial expected quality in Table 2. The inclusion task was more difficult for low-quality clusters such as the *Random* cluster. In contrast, for the high-quality clusters *NFL*, *Macedonia*, and *Oil*, participants found it easier to discover patterns. Notably, the medium-quality cluster *Astronomy* gets a high score which was due to it also containing contrasting articles similar to *Macedonia*. We concluded that human interpretation of cluster coherency can be most effectively quantified with data from the Likert scale questions asked to the participants after completing the inclusion and naming task.

4.1.4 Interpretation and Agreement Score IA

The interpretation and agreement score, IA, summarizes the agreement and complexity estimation metrics designed to indicate when participant interpretation of a cluster varies. In contrast to CP, it highlighted that the participants to a lesser extent shared a unified interpretation of the medium quality clusters *Financial Advice* and *Tips & Tricks* prompting the investigator to do further analysis. Another example is the high-quality cluster *South Africa* that scored lower than anticipated. Inspecting the free-text naming in Table 3 revealed that the cluster contained a mix of general South African articles and articles about a specific debate on South African land issues and politics.

The artificially constructed cluster *Cannabis/IT Security* had high IA scores (Figure 4) even though the CP scores were low. We can also see that for this perfectly split cluster, the scores are largely influenced by which topic the participants chose to focus on, *Cannabis* or *IT Security*. When a clear majority has chosen one side (FT and UF), the IA score is closer to the best clusters. When they are more equally divided between the topics, like in TAX, then the agreement metrics are reduced while the complexity estimation metrics stay on similar levels. Some participants did not choose a side and instead found an umbrella name for the two topics. The *Cannabis/IT Security*, similar to the *Macedonia* and *Astronomy* clusters, highlights that a high $\mathrm{A}^{\mathrm{seg}}$ score indicates that there are two contrasting groups of articles in the sample.

In summary, the results show that CIPHE quantitative metrics provide valuable insights into cluster analysis. CIPHE successfully identified which clusters were interpreted in multiple ways, and the provided inclusion task was able to quantify cluster precision indirectly.

4.2 Instruction Comparison

The case study compares the instruction sets Free Text (FT), Unifying Feature (UF), and Taxonomy (TAX) to investigate how different instructions affect the results. As Figure 4 shows, $\mathrm{A}^{\mathrm{inc}}$ and $\mathrm{A}^{\mathrm{seg}}$ have significant influence when the IA score differs between instructions on the same cluster. However, no instruction set shows a clear pattern to affect one specific metric. This means that the differences in the instruction sets had little impact when performing the survey.

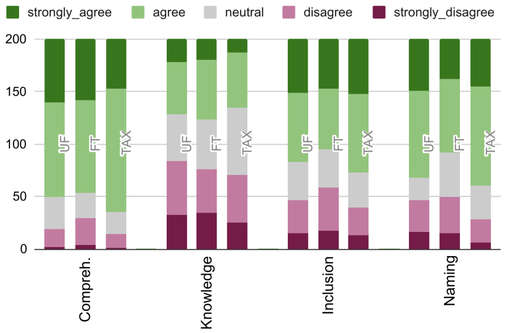


Figure 5: Summary of complexity assessments.

The *inclusion simplicity* and *naming simplicity* in Figure 5 show that the participants exposed to TAX found it slightly easier to perform the survey. The reason may be that the participants had more help with structure and vocabulary when performing the survey tasks. Given that creating a taxonomy may be labor-intensive and limits free naming, we consider FT to be a suitable instruction set for general purposes, as it is the least restrictive and no clear drawbacks were observed in the results.

4.3 Adaptions and Improvements

A CIPHE evaluation is admittedly more complicated than other annotation tasks. It requires participant attention to contextualize a set of articles and to make complex decisions through their reading comprehension and knowledge. An important aspect of the case study was confirming the feasibility of performing a CIPHE survey in a crowdsourcing environment without careful screening or extensive annotation training. There were 2 rejected participants for each set of instructions, resulting in around 10% instruction failure. Improving the instructions and the survey design may reduce this number. Further in this section, we discuss some potential improvements and adaptions for CIPHE.

The cluster precision scores showed that participants generally are inclined to exclude at least one article per cluster. This makes it difficult to reach $\mathrm{CP} = 1.0$ even for high-quality clusters such as the *NFL* cluster where many participants chose at least one article at random to exclude. One potential improvement area to get accurate cluster precision scores was revealed unintentionally with the *Macedonia* cluster. Here, 8 articles were about Greek protests against Macedonia changing its name, and 2 were about Joshua Boyle, and therefore easy to identify for the participants. This results in *Macedonia* having an almost perfect IA score and an accurate CP of 0.8. The setup for this cluster sample resembles the intrusion task used in keyword-based topic coherence metrics (Chang et al., 2009). Giving contrasting articles mixed with the sample

South Africa	NFL BLM Protest	Financial Advice	Random
South African Farmers	NFL Protest policies	About money	Passings
South African news	NFL Protests	Saving on living costs.	social media news
South Africa	NFL Players Protest	Monthly expenses	UK news - miscellaneous
mobilization of farmers	NFL protests	Financial Tips	Celebrity news
Important news about South Africa	nfl espn	Financial Advice	TV News
South Africa farmers	NFL protests	Mortgage issues	UK News
South African Farming Politics	NFL kneeling protest	Economy. UK	Celebrities deaths

Table 3: A sample of participant free-text responses in the naming task.

articles helps participants contextualize in the inclusion task. However, gamifying the task to find the intruders may divert attention from attentively exploring the cluster. E.g., completely unrelated articles introduced to the *South Africa* cluster likely complicate finding the fine-grained division of general South Africa news and the land issue. Additionally, intruding articles may change the overall context of the cluster and make it less granular. E.g., including other articles about protests to the cluster *Macedonia*, likely skews the context to be generally about protests. Using contrasting intruders likely improves CP for less granular categories, while providing only articles from the samples aids exploratory work.

One motivation behind creating CIPHE was its ability to capture semantic properties beyond themes and events. The clusters *Financial Advice* and *Tips and Tricks* most clearly exhibit such properties, which can indeed be found when inspecting the free text responses (Table 3). However, one observable drawback with this version of the CIPHE survey was that the participants would often default to a topical response such as *Money* or *Mortages* instead of considering the stylistic feature of *Advice*. To focus on specific characteristics, the investigator could separate each cluster characteristic that they are interested in. E.g., preparing characteristic-specific Likert scale and free text questions for topic, style and any other wanted characteristic. This would reduce the need for the sample texts to strongly exhibit a single characteristic and aid the participants in seeing beyond the topical content.

Practitioners interested in using CIPHE for their own evaluation are encouraged to adjust the instructions and survey layout to fit their purpose.

5 Conclusion

We have presented *Cluster Interpretation and Precision from Human Exploration* (CIPHE), a method for collecting human perception data of document clusters. CIPHE is based on the assumption that humans, when presented with a random sample of texts from a given cluster, can identify a majority feature of the texts, and also determine which texts should be excluded from the cluster. This is called the *inclusion task* and it shows promise for being an indirect measurement of cluster quality that can be used for algorithmic improvement.

The case study on ten clusters using crowdsource workers showed that participants generally saw similar coarse themes and that CIPHE highlighted when a cluster contained multiple interpretation angles. The framework is flexible enough to support a variety of research questions and practical applications. It was designed to be applicable even with only one dataset and model. Future work involves conducting larger-scale experiments with varying document styles to assess statistical properties.

Data and Code Availability

The code for the CIPHE framework can be found 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.

Acknowledgements

We thank the anonymous reviewers at both NLP4DH and TACL for their insightful comments, resulting in a more refined contribution and exciting future directions. We thank the team at Aeterna Labs for providing resources and expertise, enabling the realization of the study in their concrete real-world setting. The Ph.D. candidate Anton Eklund is funded by the Swedish Foundation for Strategic Research, project ID19-0055.

References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. Agreement is overrated: A plea for correlation to assess human evaluation reliability. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 344–354, Tokyo, Japan. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

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.

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.

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

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.

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.

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.

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.

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.

David Mimno. 2012. Computational historiography: Data mining in a century of classics journals. *J. Comput. Cult. Herit.*, 5(1).

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.

David J. Newman and Sharon Block. 2006. Probabilistic topic decomposition of an eighteenth-century american newspaper. *Journal of the American Society for Information Science and Technology*, 57(6):753–767.

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.

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.

Anna Shadrova. 2021. Topic models do not model topics: epistemological remarks and steps towards best practices. *Journal of Data Mining & Digital Humanities*, 2021.

Samer Abdulateef Waheeb, Naseer Ahmed Khan, and Xuequn Shang. 2022. Topic modeling and sentiment analysis of online education in the covid-19 era using social networks based datasets. *Electronics*, 11(5).

Hanna Megan Wallach. 2008. *Structured topic models for language*. Ph.D. thesis, University of Cambridge Cambridge, UK.

Elissa Nakajima Wickham and Emily Öhman. 2022. Hate speech, Censorship, and Freedom of Speech: The Changing Policies of Reddit. *Journal of Data Mining & Digital Humanities*, NLP4DH.

Susik Yoon, Yu Meng, Dongha Lee, and Jiawei Han. 2023. Scstory: Self-supervised and continual online story discovery. In *Proceedings of the ACM Web Conference 2023*, pages 1853–1864, New York, NY, USA. Association for Computing Machinery.

A Detailed Participant Demography

The detailed participant demography is given in Table 4.

	N	Percentage
Continent		
Europe	45	75%
Africa	9	15%
Asia & Oceania	4	7%
North America	2	3%
Age		
18-29	41	68%
30-39	10	17%
40-49	6	10%
50+	3	5%
Education		
High School	15	25%
Technical college	8	13%
Undergraduate	21	35%
Graduate	14	23%
Doctorate	2	3%
Total		
Overall	66	100%
Rejected	6	9%
Used	60	91%

Table 4: Demography of the participants.

B Survey Details

B.1 Consent

Welcome!
Thank you for participating in this study! Your input is helping us develop transparent ways of evaluating AI.

AI models can be used to organize huge amounts of text documents. To the human eye it is not always obvious which features of the texts an AI model has cared about. Hence we see a need for a practical method for humans to evaluate how AI models organize texts.

For this study such AI models have been applied to group news articles. Your role in this study is to test an evaluation tool on these groups of news articles, and to assess whether the groups make sense. You are given 10 such groups for evaluation. The articles are in English and published in 2018 and 2022.

Your answers are anonymous and will not be used as training data for models. By agreeing to participate in this study, you consent to have your anonymous responses be stored at Umeå University and included in any research paper using this data.

For further questions or comments, contact the principal investigator Anton Eklund at anton.eklund@cs.umu.se.

B.2 Instructions to participants

B.2.1 UF

Introduction The evaluation tool is based on the assumption that the texts in each group have something in common. We call this a unifying feature. Unifying features are often Themes (e.g., Football, Politics, Natural Disasters, a person, a city), or Events (e.g., a particular football match, an election, a hurricane). Sometimes, the unifying feature is something else (e.g., opinions, formality) which are more difficult to identify but may still give insight into the AI models.

Your task is to explore the group and assess if any unifying feature exists. The models may have made mistakes when grouping the texts. If so, you should remove some texts from the group and define the feature from the remaining texts. You will be given the opportunity to explain why some texts are removed.

Explore the Group:

- Read all the titles carefully.
- Click on the title to show the text body of an article if you are unsure what it is about.
- Make up your mind what is most unifying for these articles.
- Exclude articles that don't fit with the rest of the group.

Unifying Feature:

- Select one of the following unifying feature types:

  - Event: Something placed in time (e.g., a particular football match, an election, a hurricane).
  - Theme: Not bound by time (e.g., Football, Politics, Natural Disasters, a person, a city).
  - Other: Something unites the articles but is not an Event or a Theme (e.g., language style, opinion, formality).

- Name the group.

- Optionally write a comment about why you are excluding some articles.

Rate the complexity of the task:

- You are asked to rate your agreement to statements about your experience assessing the group.

B.2.2 FT

Introduction The evaluation tool is based on the assumption that the texts in each group have something in common. It could be a theme (e.g., Football, Politics, Natural Disasters, a person, a city), or an event (e.g., a particular football match, an election, a hurricane). Sometimes, it is something else (e.g., opinions, formality) which may be more difficult to identify but will still give insight into the AI models.

Your task is to explore the group and asses whether the articles have anything in common. The models may have made mistakes when grouping the texts. If so, you should remove some texts from the group and name what is common in the remaining texts.

Explore the Group:

- Read all the titles carefully.

- Click on the title to show the text body of an article if you are unsure what it is about.

- Make up your mind what is most unifying for these articles.

- Exclude articles that don't fit with the rest of the group.

Name the group:

- Write a descriptive title of the group. Examples: "Football", "Eurovision Song contest 2022", "First-person stories"

Rate the complexity of the task:

- You are asked to rate your agreement to statements about your experience assessing the group.

B.2.3 TAX

Introduction The evaluation tool is based on the assumption that the texts in each group have something in common. It could be a theme (e.g., Football, Politics, Natural Disasters, a person, a city), or an event (e.g., a particular football match, an election, a hurricane). Sometimes, it is something else (e.g., opinions, formality) which may be more difficult to identify but will still give insight into the AI models.

Your task is to explore the group and assess whether the articles have anything in common. The models may have made mistakes when grouping the texts. If so, you should remove some texts from the group and name what is common in the remaining texts.

Explore the Group:

- Read all the titles carefully.

- Click on the title to show the text body of an article if you are unsure what it is about.

- Make up your mind what is most unifying for these articles.

- Exclude articles that don't fit with the rest of the group.

Define the group:

- Select one of the following general categories that you think the group fits under:
  - Culture
  - Entertainment
  - Politics
  - Crime
  - War
  - Lifestyle
  - Science
  - Home & Garden
  - Sports
  - Business & Finance
  - Personal Finance
  - Automotive
  - Weather

  - Technology
  - Environment
  - Real Estate
  - Other

- Give a descriptive name to the group.

Rate the complexity of the task:

- You are asked to rate your agreement to statements about your experience assessing the group.