CluHTM - Semantic Hierarchical Topic Modeling based on CluWords

Felipe Viegas[1], Washington Cunha[1], Christian Gomes[1], Antônio Pereira[2], Leonardo Rocha[2], Marcos André Gonçalves[1]
[1] Universidade Federal de Minas Gerais - Brazil
[2] Universidade Federal de São João del Rei - Brazil
{frviegas, washingtoncunha}@dcc.ufmg.br
{christianreis, mgoncalv}@dcc.ufmg.br
{antoniopereira, lcrocha}@ufsj.edu.br

Abstract

Hierarchical Topic modeling (HTM) exploits latent topics and relationships among them as a powerful tool for data analysis and exploration. Despite advantages over traditional topic modeling, HTM poses its own challenges, such as (1) topic incoherence, (2) unreasonable (hierarchical) structure, and (3) issues related to the definition of the "ideal" number of topics and depth of the hierarchy. In this paper, we advance the state-of-the-art on HTM by means of the design and evaluation of *CluHTM*, a novel non-probabilistic hierarchical matrix factorization aimed at solving the specific issues of HTM. CluHTM's novel contributions include: (i) the exploration of richer text representation that encapsulates both, *global (dataset level)* and local semantic information – when combined, these pieces of information help to solve the *topic incoherence* problem as well as issues related to the *unreasonable structure*; (ii) the exploitation of a stability analysis metric for defining the number of topics and the "shape" the hierarchical structure. In our evaluation, considering twelve datasets and seven state-of-the-art baselines, CluHTM outperformed the baselines in the vast majority of the cases, with gains of around 500% over the strongest state-of-the-art baselines. We also provide qualitative and quantitative statistical analyses of why our solution works so well.

1 Introduction

Topic Modeling (TM) is the task of automatically extracting latent topics (e.g., a concept or a theme) from a collection of textual documents. Such topics are usually defined as a probability distribution over a fixed vocabulary (a set of words) that refers to some subject and describes the latent topic as a whole. Topics might be related to each other, and if they are defined at different semantic granularity levels (more general or more specific), this naturally induces a hierarchical structure. Although traditional TM strategies are of great importance to extract latent topics, the relationships among them are also extremely valuable for data analysis and exploration. In this context, Hierarchical Topic Modeling (HTM) aims to achieve – to induce latent topics from text data while preserving the inherent hierarchical structure (Teh et al., 2006). Relevant scenarios have been shown to enjoy the usefulness of HTM, such as (i) hierarchical categorization of Web pages (Ming et al., 2010), (ii) extracting aspects hierarchies in reviews (Kim et al., 2013) and (iii) discovering research topics hierarchies in academic repositories (Paisley et al., 2014).

Despite its practical importance and potential advantages over traditional TM, HTM poses its own challenges, the main ones being: (i) *topic incoherence* and (ii) *unreasonable hierarchical structure*. *Topic Incoherence* has to do with the need to learn meaningful topics. That is, the top words that represent a topic have to be semantically consistent with each other. *Unreasonable structure* is related to the extracted hierarchical topic structure. Topics near the root should be more general, while topics close to the leaves should be more specific. Furthermore, child topics must be coherent with their corresponding parent topics, guaranteeing a reasonable hierarchical structure. Finally, (iii) the number of topics in each hierarchy level is usually unknown and cannot be previously set to a predefined value since it directly depends on the latent topical distribution of the data.

Both supervised and unsupervised approaches have been applied to HTM. Supervised methods use prior knowledge to build the hierarchical tree structure, such as labeled data or linking relationships among documents (Wang et al., 2015). Those strategies are unfeasible when there is no explicit taxonomy or hierarchical scheme to associate with documents or when such an association (a.k.a., labeling) is very cumbersome or costly to obtain. Unsupervised HTM (uHTM)

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8138–8150

deals with such limitations. uHTM methods do not rely on previous knowledge (such as taxonomies or labeled hierarchies), having the additional challenge of discovering the hierarchy of topics based solely on the data at hand.

HTM solutions can also be roughly grouped into non-probabilistic and probabilistic models. In probabilistic strategies, textual data is considered to be "ruled" by an unknown probability distribution that governs the relationships between documents and topics, hierarchically. The major drawback in this type of approach has to do with the number of parameters in the model, which rapidly grows with the number of documents. This leads to learning inefficiencies and proneness to over-fitting, mainly for short textual data (Tang et al., 2014). To overcome these drawbacks, non-probabilistic models aim at extracting hierarchical topic models through matrix factorization techniques instead of learning probability distributions. Such strategies also pose challenges. They are usually limited to just local information (i.e., data limitation) as they go deeper into the hierarchy when extracting the latent topics. That is, as one moves more in-depth in the hierarchical structure representing the latent topics, the available data rapidly reduces in size, directly impacting the quality of extracted topics (in terms of both coherence and structure reasonableness). Probabilistic models mitigate this phenomenon as they rely on global information when handling the probability distributions(Xu et al., 2018). Because of that, the current main HTM methods are built based on probabilistic methods (Griffiths et al., 2004; Mimno et al., 2007).

In this paper, we aim at exploring the best properties of both non-probabilistic and probabilistic strategies while mitigating their main drawbacks. Up to our knowledge, the only work to explore this research venue is (Liu et al., 2018). In that work, the authors explore NMF for solving HTM tasks by enforcing three optimization constraints during matrix factorization: global independence, local independence, and information consistency. Those constraints allow their strategy, named HSOC, to produce hierarchical topics that somehow preserve topic coherence and reasonable hierarchical structures. However, as we shall see in our experiments, HSOC is still not capable of extracting coherent topics when applied to short text data, which is currently prominent on the Web, especially on social network environments.

We here propose a distinct approach, taking a data engineering perspective, instead of focusing on the optimization process. More specifically, we explore a matrix factorization solution properly designed to explore global information (akin to probabilistic models) when learning hierarchical topics while ensuring proper topic coherence and structure reasonableness. This strategy allows us to build a data-efficient HTM strategy, less prone to over-fitting that also enjoys the desired properties of topic coherence and reasonable (hierarchical) structure. We do so by applying a matrix factorization method over a richer text representation that encapsulates both, global and semantic information when extracting the hierarchical topics.

Recent non-probabilistic methods (Shi et al., 2018; Viegas et al., 2019) have produced top-notch results on traditional TM tasks by taking advantage of semantic similarities obtained from distances between words within an embedding space (Mikolov et al., 2013; Pennington et al., 2014). Our critical insight for HTM was to note that the richer (semantic) representation offered by distributional word embeddings can be readily explored as a global[1] source of information in more profound levels of the hierarchical structure of topics. This insight gives us an essential building block to overcome the challenges of matrix factorization strategies for HTM without the need for additional optimization constraints.

In (Viegas et al., 2019), the authors exploit the nearest words of a given "pre-trained" word embedding to generate "meta-words", *aka Cluwords*, able of expanding and enhancing the document representation in terms of syntactic and semantic information. Such an improved representation is capable of mitigating the drawbacks of using the projected space of word embeddings as well as extracting cohesive topics when applying non-negative matrix factorization for topic modeling.

Motivated by this finding, we here advance the state-of-the-art in HTM, by designing, developing and evaluating an unsupervised non-probabilistic HTM method that exploits CluWords as a key building block for TM when capturing the latent hierarchical structure of topics. We focus on the NMF method for uncovering the latent hierarchy as it is the most effective matrix factorization method for our purposes. Finally, the last aspect needed

[1]Distances in the embeddings space are global as they do consider the whole vocabulary and interactions among words in specific contexts.

to be addressed for the successful use of NMF for HTM is the definition of the appropriate number of topics *k* to be extracted. Choosing just a few topics will produce overly broad results while choosing too many will result in over-clustering the data into many redundant, highly-similar topics. Thus, our proposed method uses a stability analysis concept to automatically select the best number of topics for each level of the hierarchy.

As we shall see, our approach outperforms HSOC and hLDA (current state-of-the-art) for both small and large text datasets, often by large margins. To summarize, our main contributions are: (i) a *novel non-probabilistic HTM strategy – CluHTM – based on NMF and CluWords* that excels on HTM tasks (in both short and large text data) while ensuring topic coherence and reasonable topic hierarchies; (ii) the exploitation in an original way of a cross-level stability analysis metric for defining the number of topics and ultimately 'the shape' of the hierarchical structure; as far as we know *this metric has never been applied with this goal*; (iii) an *extensive empirical analysis* of our proposal considering twelve datasets and seven state-of-the-art baselines. In our experimental evaluation, CluHTM outperformed the baselines in the vast majority of the cases (In case of NPMI, in **all** cases), with gains of 500% when compared to hLDA and 549% when compared to HSOC, some of the strongest baselines; and finally, (iv) qualitative and quantitative statistical analyses of the individual components of our solution.

2 Related Work

Hierarchical Topic Modeling (HTM) can be roughly grouped into supervised and unsupervised methods. Considering the supervised HTM strategies, we here highlight some relevant supervised extensions to the traditional Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a widely used strategy for the topic modeling (TM). **LDA** assumes a Dirichlet probability distribution over textual data to estimate the probabilities of words for each topic. In (Mcauliffe and Blei, 2008), the authors propose **SLDA**, a supervised extension of LDA that provides a statistical model for labeled documents. **SLDA** allows connecting each document to a regression variable to find latent topics that will best predict the response variables for future unlabeled documents. Based on **SLDA**, Hierarchical Supervised LDA (HSLDA) (Perotte et al., 2011) incorporates the hierarchy of multi-label and pre-labeled data into a single model, thus providing extended prediction capabilities w.r.t., the latent hierarchical topics. The Supervised Nested LDA (SNLDA) (Resnik et al., 2015), also based on **SLDA**, implements a generative probabilistic strategy where topics are sampled from a probability distribution. **SNLDA** extends SLDA by assuming that the topics are organized into a tree structure. Although our focus is on unsupervised solutions, *we include **SLDA**, **HSLDA** and **SNLDA** as baselines in our experimental evaluation.*

We now turn our attention to unsupervised HTM strategies, in which a hierarchical structure is learned during topic extraction. In (Mimno et al., 2007) the authors propose Hierarchical Pachinko Allocation Model (hPAM), an extension of *Pachinko Allocation* (PAM) (Li and McCallum, 2006). In PAM, documents are a mix of distributions over an individual topic set, using a directed acyclic graph to represent the co-occurrences of topics. Each node in such a graph represents a *Dirichlet* distribution. At the highest level of PAM, there is only a single node, where the lowest levels represent a distribution between nodes of the next higher level. In hPAM, each node is associated with a distribution over the vocabulary of documents.

In (Griffiths et al., 2004), the authors propose the **hLDA** algorithm, which is also an expansion of LDA, being considered state-of-the-art in HTM. In hLDA, in addition to using the text Dirichlet distribution, the nested Chinese Restaurant Process (nCRP) is used to generate a hierarchical tree. NCRP needs two parameters: the tree level and a γ parameter. At each node of the tree, a document can belong to a path or create a new tree path with probability controlled by γ. More recently, in (Xu et al., 2018), the authors propose the unsupervised HTM strategy named a knowledge-based hierarchical topic model (**KHTM**). This method is based on hLDA and, as such, models a generative process whose parameter estimation strategy is based on Gibbs sampling. KHTM is able to uncover prior knowledge (such as the semantic correlation among words), organizing them into a hierarchy, consisting of knowledge sets (k-sets). More specifically, the method first generates, through hLDA, an initial set of topics. After comparing pairs of topics, those topics with similarity higher than α (a.k.a., k-sets) are then filtered so that the first 20 words of each topic are kept, and the remaining are just discarded. Those

extracted k-sets are then used as an extra weight when extracting the final topics. *All these methods are used as baselines in our experimentation.*

Probably the most similar work to ours is the HSOC strategy, proposed in (Liu et al., 2018), which proposes to use NMF for solving HTM tasks. In order to mitigate the main drawbacks of NMF in the HTM setting[2], HSOC relies on three optimization constraints to properly drive the matrix factorization operations when uncovering the hierarchical topic structure. Such constraints are global independence, local independence, and information consistency, and allow HSOC to derive hierarchical topics that somehow preserve topic coherence and reasonable hierarchical structures.

As it can be observed, almost all models, supervised or unsupervised, are based on LDA. As discussed in Section 1, though matrix factorization strategies normally present better results than Dirichlet strategies in TM tasks, for HTM, the situation is quite different. In fact, matrix factorization methods face difficult challenges in HTM, mainly regarding data size as ones go deeper into the hierarchy. More specifically, at every hierarchical level, a matrix factorization needs to be applied to increasingly smaller data sets, ultimately leading to insufficient data at lower hierarchy levels. These approaches also do not exploit semantics nor any external enrichment, relying only on the statistical information extracted from the dataset. Contrarily, here we propose a new HTM approach, called **CluHTM**, which exploits externally built *word embedding* models to incorporate *global* semantic information into the hierarchical topic tree creation. This brings some important advantages to our proposal in terms of effectiveness, topic coherence, and hierarchy reasonableness altogether.

3 Background

3.1 CluWords Representation

Cluwords (Viegas et al., 2019) combine the traditional Bag of Words (BoW) statistical representation with semantic information related to the words present in the documents. The semantic context is obtained employing a "pre-trained" word representation, such as Fasttext (Mikolov et al., 2018). Figure 1 presents the process of transforming each original word into a Cluword (cluster of words) representation. First, the strategy uses the information about the dataset, as well as pre-trained word embedding (i.e. Fasttext) to build semantic relationships between a word and its neighbors (described in Section 3.1.1). Next, statistical information on words (e.g., term frequency, document frequency) is extracted from the dataset. Then, both semantic and statistical information are combined to measure the importance of each Cluword as explained in Section 3.1.2. Cluwords enjoy the best of "two worlds": it conjugates statistical information on the dataset, which has demonstrated to be very effective, efficient and robust in text applications, enriched with semantic contextual information captured by distributional word embeddings adapted to the dataset by the clusterization process described next.

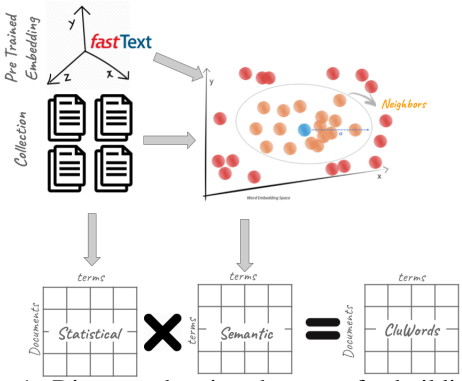

Figure 1: Diagram showing the steps for building the CluWords representation.

3.1.1 Cluwords Generation

Let $\mathcal{W}$ be the set of vectors representing each word t in the dataset vocabulary (represented as $\mathcal{V}$). Each word $t \in \mathcal{V}$ has a corresponding vector $u \in \mathcal{W}$. The CluWords representation is defined as in Figure 1. The semantic matrix in the Figure 1 is defined as $C \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$, where each dimension has the size of the vocabulary ($|\mathcal{V}|$), t' represents the rows of C while t represents the columns. Finally, each index $C_{t',t}$ is computed according to Eq. 1.

$$C_{t',t} = \begin{cases} \omega(u_{t'}, u_t) & \text{if } \omega(u_{t'}, u_t) \geq \alpha \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where $\omega(u_{t'}, u_t)$ is the cosine similarity and α is a similarity threshold that acts as a regularizer for the representation. Larger values of α lead sparser representations. In this notation each column t of the semantic matrix C will be forming a CluWord t and each value of the matrix $C_{t',t}$ may receive the cosine similarity between the vectors $u_{t'}$ and u_t in the embedding space $\mathcal{W}$ if it is greater than

[2]Namely, the incoherence of topics and unreasonable hierarchical structure caused by the lack of a learned probability distribution that governs the document/topics relationships

or equal to α . Otherwise, the $C_{t',t}$ receives zero, according to the Eq. 1.

3.1.2 TFIDF Weight for CluWords

In Figure 1, the CluWords representation is defined as the product between the statistical matrix (a.k.a. term-frequency matrix) and semantic matrix C. The statistical matrix (TF) can be represented as a $TF \in \mathbb{R}^{|\mathcal{D}|\times|\mathcal{V}|}$, where each position $TF_{d,t}$ relates to the frequency of a word t in document d. Thus, given a CluWord (CW) t for a document d, its data representation corresponds to $\text{CW}_{d,t} = \overrightarrow{TF_d} \times \overrightarrow{C_{,t}}$, where $\overrightarrow{TF_d}$ has the term-frequencies of document d, and $\overrightarrow{C_{,t}}$ is the semantic scores for the CluWord t, according to Eq. 1.

The TFIDF weighting for a CluWord t in a document d is defined as $CW_{d,t} = CW_{d,t} \times idf_t$. The IDF component is defined as $idf_t = log\left(\frac{|\mathcal{D}|}{\sum_{1\leq d\leq|\mathcal{D}|}\mu_{t,d}}\right)$, where $\mathcal{D}$ is the number of documents and $\mu_{t,d}$ is the average of semantic weights of the semantic matrix C for the CluWord t ($\overrightarrow{C_{,t}}$) that occurs in the vocabulary $\mathcal{V}_d$. The average $\mu_{t,d}$ is defined as $\mu_{t,d} = \frac{1}{|\mathcal{V}_{d,t'}|} \cdot \sum_{t'\in(\mathcal{V}_d\cap\overrightarrow{C_{,t}})} C_{t',t}$.

3.2 Stability Measure

The Stability measure is motivated by the term-centering approach generally taken in topic modeling strategies, where topics are usually summarized as a truncated set of top words (Greene et al., 2014).

The intuition behind this strategy is, given some K topics, to measure whether running multiple random samplings for a topic modeling strategy results in Stability, in terms of p top words extracted from the topics. Given a range of topics $[\mathcal{K}_{min}, \mathcal{K}_{max}]$, and some topic modeling strategy (on our case, Non-negative Factorization Matrix method), the strategy proceeds as follows. First, it learns a topic model considering the complete data set representation $\mathcal{D}$, which will be used as a reference point ($\mathcal{W}_\mathcal{D}$) for analyzing the Stability afforded by the K topics. Note that the p top words represent each topic. Subsequently, $\mathcal{S}$ samples of the data are randomly drawn from $\mathcal{D}$ without replacement, forming a subset of $\mathcal{D}'$ documents. Then, $|\mathcal{S}|$ topic models are generated, one for each subsampling ($\mathcal{W}_{\mathcal{S}_i}$).

To measure the quality of K topics, the Stability computes the mean agreement among each pair of ($\mathcal{W}_\mathcal{D}$, $\mathcal{W}_{\mathcal{S}_i}$). The goal is to find the best match between the p top words of the compared topics. The agreement is defined as $agree(\mathcal{W}_x, \mathcal{W}_y) = \frac{1}{p}\sum_{i=1}^{p} AJ(w_{xi}, \rho(w_{xi}))$, where $AJ(\cdot)$ is the average Jaccard coefficient used to compare the similarity among the words w and $\rho(\cdot)$ is the optimal permutation of the words in $\mathcal{W}_{\mathcal{S}_i}$ that can be found in $\mathbb{O}(p^3)$ time by solving the minimal weight bipartite matching problem using the Hungarian method (Kuhn, 2010).

4 Proposed Solution

CluHTM is an iterative method able to automatically define the best number of topics in each hierarchy, given a range of possible number of topics $[\mathcal{K}_{min}, \mathcal{K}_{max}]$. CluHTM explores Cluwords and Non-negative Matrix Factorization (NMF) (Lee and Seung, 2001), one of the main non-probabilistic strategies. Finally, the Stability method (described in Section 3) is used to select NMF k parameters (a.k.a number of topics).

CluHTM has five inputs (Algorithm 1), (i) $\mathcal{D}_{max}$ corresponds to the depth down to which we want to extract the hierarchical structure. (ii) $\mathcal{K}_{min}$ and $\mathcal{K}_{max}$ control the range of some topics, such range will be used in all levels of the hierarchy; (iii) $\mathcal{T}$ is the input text data; and (iv) $\mathcal{W}$ is the "pre-trained" word embedding vector space used in the CluWords generation. The output is the hierarchical structure $\mathcal{H}$ of p top words for each topic.

Algorithm 1: CluHTM

Input: $\mathcal{D}_{max}$ - Hierarchy Depth;
$\mathcal{K}_{min}$ - Number of minimum topics;
$\mathcal{K}_{max}$ - Number of maximum topics;
$\mathcal{T}$ - Term-frequency representation;
$\mathcal{W}$ - Word embedding vectors $\in \mathcal{T}$;
Output: $\mathcal{H}$ - Hierarchical Structure.

```
1  parent ← −1;
2  queue.push(0, T);
3  while queue ≠ ∅ do
4      depth, T' ← queue.pop();
5      Clu ← GenerateCluwords(T', W);
6      K ← Stability(K_min, K_max, Clu)
7      O ← NMF(Clu, K)
8      topics ← ExtractTopics(O)
9      foreach topic ∈ topics do
10         parent ← parent ∪ topic;
11         H ← H ∪ topic;
12         if depth + 1 ≤ D_max then
13             T' ← ExtractDocs(topic);
14             queue.push(depth + 1, T')
15 return H
```

The method starts by getting the root topic (line 2-3 of Algorithm 1), which is composed of all documents in $\mathcal{T}$. Since the method is iterative, each iteration is controlled by a queue schema to build a hierarchical structure. Thus, at each iteration (line 3), the algorithm produces the

CluWords representation for the documents $\in \mathcal{T}'$ (line 5), chooses the number of topics, exploiting the Stability measure (line 6), and runs the NMF method (line 7) to extract the p words for each topic in $\mathcal{O}$ (line 8). Then, in the loop of line 9, each topic is stored in the queue, as well as the respective documents of each topic.

Summarizing, our solution exploits *global* semantic information (captured by CluWords) within *local* factorizations, limited by a stability criterion that defines the 'shape' of the hierarchical structure. Though simple (and original), the combination of these ideas is extremely powerful for solving the HTM task, as we will see next.

5 Experimental Results

5.1 Experimental Setup

The primary goal of our solution is to effectively perform hierarchical topic modeling so that more coherent topics can be extracted. To evaluate topic model coherence, we consider 12 real-world datasets as reference. All of them were obtained from previous works in the literature. For all datasets, we performed stopwords removal (using the standard SMART list) and removed words such as adverbs, using the VADER lexicon dictionary (Hutto and Gilbert, 2014), as the vast majority of the essential words for identifying topics are nouns and verbs. These procedures improved both the efficiency and effectiveness of all analyzed strategies. Table 1 provides a summary of the reference datasets, reporting the number of features (words) and documents, as well as the mean number of words per document (density) and the corresponding references.

Table 1: Dataset characteristics

Dataset	#Feat	#Doc	Density
Angrybirds	1,903	1,428	7.135
Dropbox	2,430	1,909	9.501
Evernote	6,307	8,273	11.002
InfoVis-Vast [3]	6,104	909	86.215
Pinterest	2,174	3,168	4.478
TripAdvisor	3,152	2,816	8.532
Tweets	8,029	12,030	4.450
WhatsApp	1,777	2,956	3.103
20NewsGroup [4]	29,842	15,411	76.408
ACM	16,811	22,384	30.428
Uber	5,517	11,541	7.868
Facebook	5,168	12,297	6.427

[3]https://www.cc.gatech.edu/gvu/ii/jigsaw/datafiles.html
[4]http://qwone.com/~jason/20Newsgroups/

We compare the HTM strategies using representative topic quality metrics in the literature (Nikolenko, 2016; Nikolenko et al., 2017). We consider three classes of topic quality metrics based on three criteria: (a) coherence, (b) mutual information, and (c) semantic representation. In this paper, we focus on these three criteria since they are the most used metrics in the literature (Shi et al., 2018). We consider three topic lengths (5, 10 and 20 words) for each parameter in our evaluation, since different lengths may bring different challenges.

Regarding the metrics, *coherence* captures easiness of interpretation by co-occurrence. Words that frequently co-occur in similar contexts in a corpus are easier to correlate since they usually define a more well-defined "concept" or "topic". We employ an improved version of regular coherence (Nikolenko, 2016), called Coherence, defined as

$$c(t, W_t) = \sum_{w_1, w_2 \in W_t} log \frac{d(w_1, w_2) + \varepsilon}{d(w_1)}, \quad (2)$$

where $d(w1)$ denotes the number of occurrences of $w1$, $d(w1, w2)$ is the number of documents that contain both $w1$ and $w2$ together, and ε is a smoothing factor used for preventing $log(0)$.

Another class of topic quality metrics is based on the notion of *pairwise pointwise mutual information (PMI)* between the top words in a topic. It captures how much one "gains" in the information given the occurrence of the other word, taking dependencies between words into consideration. Following a recent work (Nikolenko, 2016), we here compute a *normalized version of PMI* (NPMI) where, for a given ordered set of top words $W_t = (w_1, ..., w_N)$ in a topic:

$$NPMI_t = \sum_{i<j} \frac{log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}}{-logp(w_i, w_j)}. \quad (3)$$

Finally, the third class of metrics is based on the distributed word representations introduced in (Nikolenko, 2016). The intuition is that, in a well-defined topic, the words should be semantically similar, or at least related, to be easily interpreted by humans. In a d-dimensional vector space model in which every vocabulary word $w \in W$ has been assigned to a vector $v_w \in R^d$, the vectors corresponding to the top words in a topic should be close to each other. In (Nikolenko, 2016), the authors define topic quality as the average distance between the top words in the topic, as follows:

$$W2V - L1 = \frac{1}{|W_t|(|W_t|-1)} \sum_{w_1 \neq w_2 \in W_t} d_{cos}(v_{w_1}, v_{w_2}). \quad (4)$$

Generally speaking, let $d(w_1, w_2)$ be a distance function in R^d. In this case, larger $d(w_1, w_2)$ corresponds to worse topics (with words not as localized as in topics with smaller average distances). In (Nikolenko, 2016), the authors suggest four different distance metrics, with cosine distance achieving the best results. We here also employ the cosine distance, defined as $d_{cos}(x, y) = 1 - x^T y$.

We compare our approach described in Section 4, with seven hierarchical topic model strategies marked in bold in Section 2. For the input parameters of CluHTM (Algorithm 1), we set $\mathcal{K}_{min} = 5$, $\mathcal{K}_{max}$=25, $\mathcal{R} = 10$ and $\mathcal{D}_{max} = 3$. We define $\mathcal{K}_{min}$ through empirical experiments, and the $\mathcal{K}_{max}$ was defined according to the number of topics exploited in (Viegas et al., 2019). For the baseline methods, we adopt the parameters suggested by their own works. We assess the statistical significance of our results employing a paired t-test with 95% confidence and Holm-Bonferroni correction to account for multiple tests.

5.2 Experimental Results

We start by comparing CluHTM against four state-of-the-art uHTM baselines considering the twelve reference datasets. Three hierarchical levels for each strategy are used in this comparison. In Figures 2, 4 and 3 we contrast the results of our proposed CluHTM and the reference strategies, considering the NPMI, W2V-L1, and Coherence metrics.

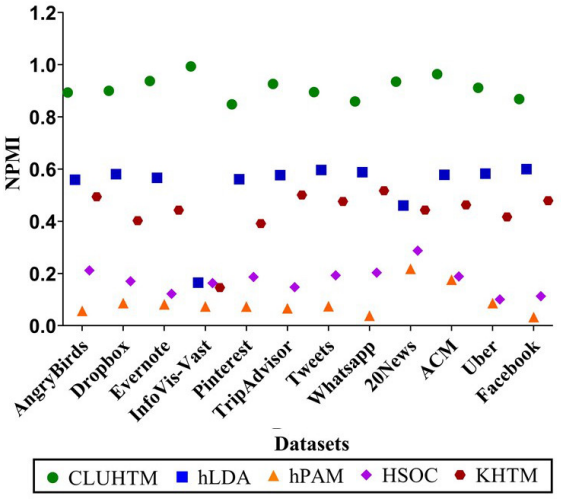

Figure 2: uHTM Comparative Results (NPMI).

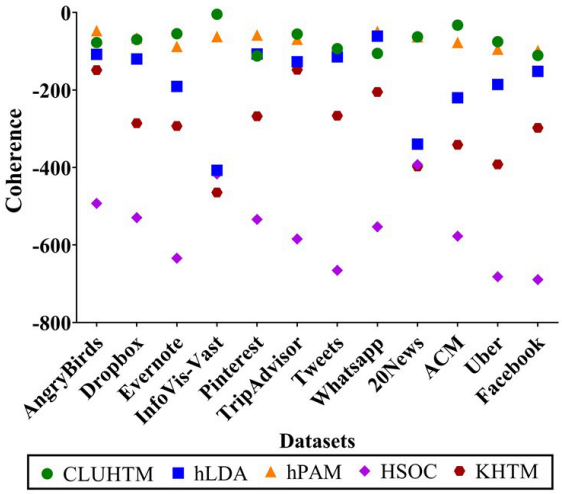

Figure 3: uHTM Comparative Results (Coherence)

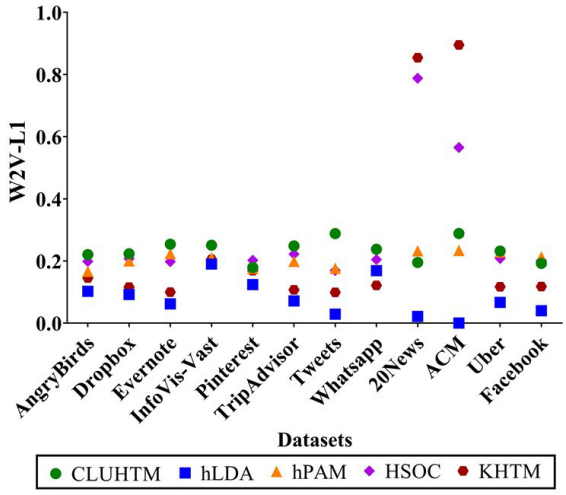

Figure 4: uHTM Comparative Results (W2V-L1)

Note that each strategy extracted a different number of topics in its hierarchical structure. Considering NPMI, the most important metric to evaluate the quality of topics (Nikolenko, 2016), we can see in Figure 2 that our strategy outperforms all baselines in all datasets by large margins, with gains over 500% against some of the strongest ones. Some of these results are the highest in terms of NMPI ever reported for several of these datasets. Considering the Coherence scores (Figure 3), our strategy achieves the single best results in 2 out of 12 datasets, with gains up to 58% and 92% against the most robust baseline (hPAM), tying in 8 out 12 and losing two times for hLDA and hPAM. Similar results can be observed for the W2V-L1 metric (Figure 4) – CluHTM ties in 10 out of 12 results, with one win and one loss for KHTM. As we will see, even with very few losses in these metrics, our method proves to be

Dataset	CluHTM	SLDA	SNLDA	HSLDA
		Coherence		
20News	-62.6898 ± 21.0606 ▲	-403.3413 ± 90.2313	-410.0020 ± 71.2366	-309.9041 ± 132.5511
ACM	-32.3371 ± 29.5853 ▲	-539.6660 ± 115.2125	-507.4476 ± 108.6966	-486.4835 ± 104.9369
		W2V-L1		
20News	1.1863 ± 0.1176 ▼	0.3093 ± 0.2006	0.3456 ± 0.2051	0.0952 ± 0.1094
ACM	1.0489 ± 0.6506 ●	0.6347 ± 0.2617	0.6803 ± 0.2243	0.2816 ± 0.1567
		NPMI		
20News	0.9351 ± 0.0365 ▲	0.2714 ± 0.1157	0.2205 ± 0.0752	0.4383 ± 0.2162
ACM	0.9641 ± 0.0416 ▲	0.2071 ± 0.0579	0.2064 ± 0.0529	0.2761 ± 0.0978

Table 2: Comparing the results achieved by each supervised HTM strategy for Coherence, W2V-L1 and NPMI.

Table 3: Number of times each strategy was the top performer. CluHTM is the best performer in most cases.

Method	Metric			$\sum$
	NPMI	W2V-L1	Coherence	(Sum)
CluHTM	**12**	**11**	**10**	**33**
hPAM	0	9	8	17
hLDA	0	2	9	11
HSOC	0	9	0	9
KHTM	0	6	2	8
SNLDA	0	2	0	2
HSLDA	0	2	0	2
textbfSLDA	0	1	0	1

more consistent than the baselines.

We now turn our attention to the effectiveness of our proposal when compared to the supervised HTM strategies. We consider the 20News and ACM datasets for which have a ground truth for supervised strategies. Table 2 presents the results considering Coherence, W2V-L1, and NPMI. The statistical significance tests ensure that the best results, marked in ▲, are superior to others. The statistically equivalent results are marked in ● while statistically significant losses are marked in ▼. Once again, in Table 2, our proposed strategy achieves the best results in 4 out of 6 cases, tying with SNLDA and HSLDA in ACM and loosing only to SLDA in 20News, both considering the W2V-L1 metric. It is important to remind that, differently from these supervised baselines, our method does not use any privileged class information to build the hierarchical structure nor to extract topics.

We provide a comparative table with all experimental results[5], including the results for each extracted level of the hierarchical structure. We summarize our findings regarding the behavior of all analyzed strategies in the 12 datasets, counting the number of times each strategy figured out as a top performer[6]. The summarized results can be seen in Table 3. Our proposal is in considerable advantage over the other explored baselines, being the strategy of choice in the vast majority of cases. Overall, considering a universe of 36 experimental results (the combination of 3 evaluation metrics over 12 datasets), we obtained the best results (33 best performances), with the most robust baseline – hPAM – coming far away, with just 17 top performances. Another interesting observation is that, in terms of NPMI, CluHTM wins in **all** cases. Details of this analysis are summarized in the Appendix.

5.3 Impact of the Factors

One important open question remains to be answered: To what extent the characteristics of the dataset impact the quality of the topics generated by our strategy? To answer this question, we provide a quantitative analysis regarding the hierarchical topic modeling effectiveness, measured by the NPMI score.

We start our analysis by quantifying the effects of the parameters of interest (i.e., factors). Those factors might affect the performance of the system under study, while also determining whether the observed variations are due to significant effects (e.g., measurement errors, the inherent variability of the process being analyzed (Jain, 1991)). To this end, we adopt a *full factorial design*, which uses all the possible combinations of the levels of the factors in each complete experiment. The first factor is the dataset. The idea is to analyze the impact of textual properties such as dataset size, density, dimensionality, etc. Thus, each level of this factor is a dataset in Table 1. The second factor is the HTM strategies evaluated in the previous Section. In this factor, we intend to assess the impact of the extracted topics, as well as the hierarchical structure. Each level of this factor is an evaluated HTM strategy. All the possible combination between these two factors will be measured by the average of NMPI among topics of the hierarchical structure.

Results are shown in Table 4. In the Table, we highlight the average NPMI and the effects of each

[5] see Appendix, Section Supplementary Results for detailed results

[6] If two approaches are statistically tied as top performers in the same dataset, both will be counted.

Dataset-Algorithm	CluHTM	hLDA	hPAM	HSOC	KHTM	Row Sum	Row Mean	Row Effect
Angry Birds	0.8934	0.5593	0.3604	0.2120	0.4940	2.5191	0.5038	0.0507
Dropbox	0.9002	0.5806	0.2529	0.1703	0.4022	2.3062	0.4612	0.0082
Evernote	0.9374	0.5668	0.1534	0.1222	0.4426	2.2224	0.4445	-0.0086
Facebook	0.8686	0.5998	0.1517	0.1128	0.4791	2.2120	0.4424	-0.0107
InfoVis-Vast	0.9935	0.1650	0.1191	0.1632	0.1459	1.5867	0.3173	-0.1357
Pinterest	0.8482	0.5614	0.3028	0.1865	0.3912	2.2901	0.4580	0.0049
Trip Advisor	0.9265	0.5769	0.2745	0.1477	0.5007	2.4263	0.4853	0.0322
Tweets	0.8950	0.5966	0.2130	0.1928	0.4759	2.3733	0.4747	0.0216
Uber	0.9116	0.5829	0.1403	0.1006	0.4168	2.1522	0.4304	-0.0226
Whatsapp	0.8594	0.5881	0.3976	0.2031	0.5172	2.5654	0.5131	0.0600
Col Sum	9.0338	5.3774	2.3657	1.6112	4.2656	22.6537	-	-
Col Mean	0.9034	0.5377	0.2366	0.1611	0.4266	-	0.4531	-
Col effect	0.4503	0.0847	-0.2165	-0.2920	-0.0265	-	-	-

Table 4: Overview of the factorial desgin

Component	Sum of Degrees	% Variation	Degrees of Freedom	Mean Square	F-Computed	F-Table (0.99)
y	14.0670	-	50	-	-	-
$y_{..}$	10.2638	-	1	-	-	-
$y - y_{..}$	3.8032	100.00%	49	-	-	-
A	3.4276	90.12%	4	0.8569	127.9197	3.8903
B	0.1345	3.92%	9	0.0149	2.2303	2.9461
e	0.2412	6.34%	36	0.0067	-	-

Table 5: ANOVA Test with 99% confidence to measure the impact of each factor.

factor. From the effects, we can observe that the CluHTM impact in the NPMI value is 99.38% higher than the overall average. We can also see that hLDA has an NPMI score higher than the overall average (18.67%) and HSOC has an NPMI score of approximately 64.44% *smaller* than overall NMPI. Concerning the datasets' effects, the full factorial design experiment tells us that they have a small impact on the variation concerning the obtained average NPMI scores. We can also observe that the dataset with the most variation of NPMI is InfoVis-Vast, with a score of 29.97% smaller than the overall NPMI.

We perform a ANOVA test to assess whether the studied factors are indeed statistically significant and conclude, with 99% confidence according to the F-test, that the choice of algorithm (factor B) explains approximately 90% of the obtained NPMI values. We can also conclude that the investigated properties of the textual data (factor A), as well as the experimental errors, have a small influence on the experimental results. Summarizing, we can conclude that the characteristics of the datasets have a lower impact on the results and that the impact of CluHTM is consistent across all of them. The ANOVA test details are presented in Table 5.

6 Conclusion

We advanced the state-of-the-art in hierarchical topic modeling (HTM) by designing, implementing and evaluation a novel unsupervised non-probabilistic method – CluHTM. Our new method exploits a more elaborate (global) semantic data representation – CluWords – as well as an original application of a stability measure to define the "shape" of the hierarchy. CLUHTM excelled in terms of effectiveness, being around two times more effective than the strongest state-of-the-art baselines, considering all tested datasets and evaluation metrics. The overall gains over some of these strongest baselines are higher than 500% in some datasets. We also showed that CluHTM results are consistent across most datasets, independently of the data characteristics and idiosyncrasies. As future work, we intend to apply CluHTM in other representative applications on the Web, such as hierarchical classification by devising a supervised version of CluHTM. We also intend to incorporate some type of attention mechanism into our methods to better understand which Cluwords are more important to define certain topics.

7 Acknowledgments

This work is partially supported by CAPES, CNPq, Finep, Fapemig, Mundiale, Astrein, projects InWeb and MASWeb.

References

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

Derek Greene, Derek O'Callaghan, and Pádraig Cunningham. 2014. How many topics? stability analysis for topic models. *CoRR*.

Thomas L Griffiths, Michael I Jordan, Joshua B Tenenbaum, and David M Blei. 2004. Hierarchical topic

models and the nested chinese restaurant process. In *Advances in neural information processing systems*, pages 17–24.

Clayton J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM'14*.

Raj Jain. 1991. *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*.

Suin Kim, Jianwen Zhang, Zheng Chen, Alice Oh, and Shixia Liu. 2013. A hierarchical aspect-sentiment model for online reviews. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Harold W. Kuhn. 2010. The hungarian method for the assignment problem. In *50 Years of Integer Programming*.

Daniel D Lee and H Sebastian Seung. 2001. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*.

Wei Li and Andrew McCallum. 2006. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, pages 577–584. ACM.

Rui Liu, Xingguang Wang, Deqing Wang, Yuan Zuo, He Zhang, and Xianzhu Zheng. 2018. Topic splitting: a hierarchical topic model based on non-negative matrix factorization. *Journal of Systems Science and Systems Engineering*, 27(4):479–496.

Jon D Mcauliffe and David M Blei. 2008. Supervised topic models. In *Advances in neural information processing systems*, pages 121–128.

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

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *LREC'18*.

David Mimno, Wei Li, and Andrew McCallum. 2007. Mixtures of hierarchical topics with pachinko allocation. In *Proceedings of the 24th ICML*, pages 633–640. ACM.

Zhao-Yan Ming, Kai Wang, and Tat-Seng Chua. 2010. Prototype hierarchy based clustering for the categorization and navigation of web collections. In *Proceedings of the 33rd ACM SIGIR*, pages 2–9. ACM.

Sergey I Nikolenko. 2016. Topic quality metrics based on distributed word representations. In *SIGIR'16*.

Sergey I Nikolenko, Sergei Koltcov, and Olessia Koltsova. 2017. Topic modelling for qualitative studies. *Journal of Information Science*.

John Paisley, Chong Wang, David M Blei, and Michael I Jordan. 2014. Nested hierarchical dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):256–270.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Adler J Perotte, Frank Wood, Noemie Elhadad, and Nicholas Bartlett. 2011. Hierarchically supervised latent dirichlet allocation. In *Advances in neural information processing systems*, pages 2609–2617.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 2015. Beyond lda: exploring supervised topic modeling for depression-related language in twitter. In *Proc. of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 99–107.

Tian Shi, Kyeongpil Kang, Jaegul Choo, and Chandan K. Reddy. 2018. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *WWW '18*, pages 1105–1114.

Jian Tang, Zhaoshi Meng, XuanLong Nguyen, Qiaozhu Mei, and Ming Zhang. 2014. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of the 31st ICML'14*, pages I–190–I–198. JMLR.org.

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

Felipe Viegas, Sérgio Canuto, Christian Gomes, Washington Luiz, Thierson Rosa, Sabir Ribas, Leonardo Rocha, and Marcos André Gonçalves. 2019. Clusterwords: Exploiting semantic word clustering representation for enhanced topic modeling. In *Proceedings of WSDM '19*, pages 753–761.

Chi Wang, Jialu Liu, Nihit Desai, Marina Danilevsky, and Jiawei Han. 2015. Constructing topical hierarchies in heterogeneous information networks. *Knowledge and Information Systems*, 44(3):529–558.

Yueshen Xu, Jianwei Yin, Jianbin Huang, and Yuyu Yin. 2018. Hierarchical topic modeling with automatic knowledge mining. *Expert Systems with Applications*, 103:106–117.

A Appendix

Supplementary Results

The Tables below expand on the results of Section 5.

Datasets	CluNMF	SLDA	SNLDA	HSLDA
20News	**-62.68 ± 21.06**	-403.34 ± 90.23	-410.00 ± 71.23	-309.90 ± 132.55
ACM	**-32.33 ± 29.58**	-539.66 ± 115.21	-507.44 ± 108.69	-486.48 ± 104.93

Table 6: Overall Coherence results compared with supervised HTM strategies.

Datasets	CluNMF	SLDA	SNLDA	HSLDA
20News	**0.9351 ± 0.0365**	0.2714 ± 0.1157	0.2205 ± 0.0752	0.4383 ± 0.2162
ACM	**0.9641 ± 0.0416**	0.2071 ± 0.0579	0.2064 ± 0.0529	0.2761 ± 0.0978

Table 7: Overall NPMI results compared with supervised HTM strategies.

Datasets	CluNMF	SLDA	SNLDA	HSLDA
20News	1.1863 ± 0.1176	**0.3093 ± 0.2006**	**0.3456 ± 0.2051**	**0.0952 ± 0.1094**
ACM	**1.0489 ± 0.6506**	0.6347 ± 0.2617	**0.6803 ± 0.2243**	**0.2816 ± 0.1567**

Table 8: Overall W2V-L1 results compared with supervised HTM strategies.

Datasets	Level	CluNMF	SLDA	SNLDA	HSLDA
20News	1 Level	-12.65 ± 0.00	-403.34 ± 90.23	-428.59 ± 0.00	-317.08 ± 0.00
	2 Level	-45.37 ± 22.72	-	-426.88 ± 103.53	-194.91 ± 0.00
	3 Level	-39.03 ± 22.94	-	-403.75 ± 56.71	-313.75 ± 135.41
ACM	1 Level	-34.15 ± 16.73	-539.66 ± 115.21	-451.24 ± 0.00	-594.07 ± 0.00
	2 Level	-34.15 ± 16.7377	-	-431.38 ± 55.64	-467.37 ± 0.00
	3 Level	-27.26 ± 6.48	-	-516.92 ± 110.93	-483.32 ± 106.59

Table 9: Coherence results by level of hierarchy compared with supervised HTM strategies.

Datasets	CluNMF	hLDA	hPAM	HSOC	KHTM
20News	**-62.68 ± 21.06**	-339.45 ± 186.20	**-62.56 ± 9.81**	-393.40 ± 76.70	-397.19 ± 143.59
ACM	**-32.33 ± 29.58**	-219.85 ± 159.25	-77.20 ± 9.92	-577.08 ± 102.33	-341.39 ± 98.87
AngryBirds	**-77.39 ± 41.17**	**-107.70 ± 65.08**	**-46.80 ± 16.44**	-492.40 ± 33.12	**-148.41 ± 94.66**
Dropbox	**-69.56 ± 33.86**	**-119.76 ± 103.54**	**-65.52 ± 12.96**	-529.34 ± 28.97	-285.62 ± 88.38
Evernote	**-54.45 ± 32.81**	**-190.48 ± 123.60**	-87.33 ± 8.80	-634.04 ± 51.01	-292.90 ± 93.97
Facebook	**-110.57 ± 51.62**	**-151.63 ± 130.89**	**-98.11 ± 12.53**	-689.19 ± 51.92	-297.75 ± 96.46
InfoVis-Vast	**-4.46 ± 11.36**	-407.05 ± 74.89	-61.9381 ± 10.1260	-416.49 ± 42.25	-464.38 ± 60.96
Pinterest	-111.90 ± 48.77	**-106.95 ± 87.06**	**-58.00 ± 16.46**	-533.79 ± 42.19	-267.71 ± 84.46
TripAdvisor	**-55.60 ± 27.54**	**-126.96 ± 90.43**	**-68.95 ± 15.60**	-584.30 ± 29.14	**-147.53 ± 97.18**
Tweets	**-93.11 ± 31.38**	**-114.24 ± 71.03**	**-92.52 ± 9.71**	-665.23 ± 59.39	-266.22 ± 77.54
Uber	**-75.25 ± 39.94**	**-185.50 ± 136.32**	-94.72 ± 9.24	-681.78 ± 55.40	-391.70 ± 108.40
Whatsapp	-105.28 ± 38.64	**-60.81 ± 62.04**	**-48.48 ± 15.55**	-552.77 ± 45.50	-204.83 ± 87.73

Table 10: Overall Coherence results compared with uHTM strategies.

Datasets	CluNMF	hLDA	hPAM	HSOC	KHTM
20News	**0.9351 ± 0.0365**	0.4603 ± 0.1498	0.2176 ± 0.0622	0.2875 ± 0.0782	0.4433 ± 0.1223
ACM	**0.9641 ± 0.0416**	0.5781 ± 0.1021	0.1758 ± 0.0432	0.1889 ± 0.0490	0.4631 ± 0.0769
AngryBirds	**0.8934 ± 0.0514**	0.5593 ± 0.0565	0.3604 ± 0.1005	0.2120 ± 0.0306	0.4940 ± 0.0711
Dropbox	**0.9002 ± 0.0454**	0.5806 ± 0.0864	0.2529 ± 0.0877	0.1703 ± 0.0325	0.4022 ± 0.0615
Evernote	**0.9374 ± 0.0334**	0.5668 ± 0.0819	0.1534 ± 0.0564	0.1222 ± 0.0232	0.4426 ± 0.0763
Facebook	**0.8686 ± 0.0531**	0.5998 ± 0.0734	0.1517 ± 0.0765	0.1128 ± 0.0458	0.4791 ± 0.0744
InfoVis-Vast	**0.9935 ± 0.0190**	0.1650 ± 0.0732	0.1191 ± 0.0533	0.1632 ± 0.0504	0.1459 ± 0.0793
Pinterest	**0.8482 ± 0.0535**	0.5614 ± 0.0664	0.3028 ± 0.0988	0.1865 ± 0.0414	0.3912 ± 0.0559
TripAdvisor	**0.9265 ± 0.0344**	0.5769 ± 0.0745	0.2745 ± 0.0906	0.1477 ± 0.0306	0.5007 ± 0.0744
Tweets	**0.8950 ± 0.0323**	0.5966 ± 0.0381	0.2130 ± 0.0453	0.1928 ± 0.0534	0.4759 ± 0.0472
Uber	**0.9116 ± 0.0424**	0.5829 ± 0.0863	0.1403 ± 0.0582	0.1006 ± 0.0305	0.4168 ± 0.0861
Whatsapp	**0.8594 ± 0.0456**	0.5881 ± 0.0326	0.3976 ± 0.0750	0.2031 ± 0.0385	0.5172 ± 0.0634

Table 11: Overall NPMI results compared with uHTM strategies.

Datasets	CluNMF	hLDA	hPAM	HSOC	KHTM
20News	1.1863 ± 0.1176	1.4423 ± 0.1412	1.1318 ± 0.0860	0.3201 ± 0.2085	**0.2153 ± 0.1757**
ACM	**1.0489 ± 0.6506**	1.4741 ± 0.0915	1.1296 ± 0.0987	0.6408 ± 0.2712	**0.1544 ± 0.1522**
AngryBirds	**1.1489 ± 0.1157**	1.3236 ± 0.0327	**1.2286 ± 0.0779**	**1.1816 ± 0.0388**	**1.2603 ± 0.0285**
Dropbox	**1.1454 ± 0.0918**	1.3388 ± 0.0402	**1.1794 ± 0.0873**	**1.1687 ± 0.0417**	1.3032 ± 0.0421
Evernote	**1.0999 ± 0.1247**	1.3828 ± 0.0524	**1.1447 ± 0.0643**	**1.1825 ± 0.0453**	1.3272 ± 0.0525
Facebook	**1.1909 ± 0.1224**	1.4152 ± 0.0411	**1.1598 ± 0.0541**	**1.1767 ± 0.0561**	**1.3008 ± 0.0390**
InfoVis-Vast	**1.1047 ± 0.0867**	1.1939 ± 0.0717	**1.1651 ± 0.0651**	**1.1919 ± 0.0509**	**1.1720 ± 0.0510**
Pinterest	**1.2101 ± 0.0963**	**1.2912 ± 0.0263**	**1.2147 ± 0.0712**	**1.1760 ± 0.0495**	**1.2255 ± 0.0257**
TripAdvisor	**1.1081 ± 0.1082**	1.3686 ± 0.0464	**1.1814 ± 0.0685**	**1.1470 ± 0.0318**	1.3161 ± 0.0327
Tweets	**1.0493 ± 0.1086**	1.4315 ± 0.0314	1.2142 ± 0.0654	1.2242 ± 0.0687	1.3285 ± 0.0372
Uber	**1.1323 ± 0.1328**	1.3758 ± 0.0419	**1.1370 ± 0.0664**	**1.1677 ± 0.0381**	1.3018 ± 0.0518
Whatsapp	**1.1239 ± 0.1087**	**1.2254 ± 0.0141**	**1.2162 ± 0.0656**	**1.1732 ± 0.0423**	1.2952 ± 0.0268

Table 12: Overall W2V-L1 results compared with uHTM strategies.

Datasets	Level	CluNMF	SLDA	SNLDA	HSLDA
20News	1 Level	0.9863 ± 0.0000	0.2714 ± 0.1157	0.1829 ± 0.0000	0.6622 ± 0.0000
	2 Level	0.9386 ± 0.0311	-	0.1753 ± 0.0644	0.5728 ± 0.0000
	3 Level	0.9495 ± 0.0319	-	0.2368 ± 0.0732	0.4255 ± 0.2179
ACM	1 Level	0.9552 ± 0.0185	0.2071 ± 0.0579	0.1107 ± 0.0000	0.3060 ± 0.0000
	2 Level	0.9552 ± 0.0185	-	0.1472 ± 0.0319	0.3909 ± 0.0000
	3 Level	0.9682 ± 0.0071	-	0.2155 ± 0.0483	0.2709 ± 0.0986

Table 13: NPMI results by level of hierarchy compared with supervised HTM strategies.

Datasets	Level	CluNMF	SLDA	SNLDA	HSLDA
20News	1 Level	1.0232 ± 0.0000	0.3093 ± 0.2006	0.2961 ± 0.0000	***
	2 Level	1.1925 ± 0.1096	-	0.2625 ± 0.1157	0.3296 ± 0.0000
	3 Level	1.2060 ± 0.1183	-	0.3750 ± 0.2231	0.0902 ± 0.1025
ACM	1 Level	1.0955 ± 0.1047	0.6347 ± 0.2617	0.5365 ± 0.0000	0.1488 ± 0.0000
	2 Level	1.0955 ± 0.1047	-	0.5060 ± 0.0302	0.0074 ± 0.0000
	3 Level	1.0320 ± 0.0716	-	0.7025 ± 0.2296	0.2961 ± 0.1510

Table 14: W2V-L1 results by level of hierarchy compared with supervised HTM strategies.

Datasets	Level	CluNMF	hLDA	hPAM	HSOC	KHTM
20News	1 Level	-12.6556 ± 0.00	-323.30 ± 0.00	-64.17 ± 0.00	-389.41 ± 69.27	-334.20 ± 0.00
	2 Level	-45.37 ± 22.72	-448.93 ± 166.05	-57.03 ± 8.52	-395.88 ± 88.74	-595.39 ± 139.97
	3 Level	-39.03 ± 22.94	-328.58 ± 184.81	-63.10 ± 9.81	-394.92 ± 72.10	-389.23 ± 137.96
ACM	1 Level	-34.15 ± 16.73	-368.20 ± 0.00	-69.87 ± 0.00	-529.64 ± 107.93	-371.89 ± 0.00
	2 Level	-34.15 ± 16.73	-544.64 ± 111.50	-80.91 ± 9.70	-566.92 ± 103.18	-708.12 ± 146.29
	3 Level	-27.26 ± 6.48	-210.03 ± 149.94	-76.90 ± 9.89	-594.03 ± 95.85	-336.04 ± 87.40
AngryBirds	1 Level	-20.27 ± 0.00	-514.71 ± 0.00	-68.27 ± 0.00	-528.23 ± 17.38	-546.60 ± 0.00
	2 Level	-40.73 ± 16.47	-168.89 ± 66.22	-14.35 ± 10.30	-510.20 ± 22.09	-549.78 ± 151.09
	3 Level	-80.55 ± 40.95	-98.05 ± 57.18	-49.83 ± 13.05	-474.54 ± 28.33	-133.16 ± 46.61
Dropbox	1 Level	-12.78 ± 0.00	-487.63 ± 0.00	-65.27 ± 0.00	-536.63 ± 34.81	-527.14 ± 0.00
	2 Level	-60.01 ± 25.01	-247.70 ± 81.57	-42.53 ± 18.08	-537.71 ± 23.30	-569.05 ± 36.06
	3 Level	-70.81 ± 34.32	-100.89 ± 91.65	-67.82 ± 9.78	-523.33 ± 28.47	-270.74 ± 61.27
Evernote	1 Level	-20.85 ± 0.00	-489.14 ± 0.00	-91.59 ± 0.00	-608.34 ± 44.38	-513.78 ± 0.00
	2 Level	-29.64 ± 7.40	-364.60 ± 106.17	-76.91 ± 12.32	-620.48 ± 50.72	-634.88 ± 28.08
	3 Level	-57.31 ± 33.23	-177.60 ± 114.68	-88.33 ± 7.67	-647.25 ± 48.42	-286.84 ± 83.09
Facebook	1 Level	-57.62 ± 16.87	-589.33 ± 0.00	-85.66 ± 0.00	-663.14 ± 53.09	-607.85 ± 0.00
	2 Level	-82.49 ± 39.87	-547.12 ± 113.61	-84.63 ± 18.77	-684.94 ± 55.76	-748.77 ± 13.93
	3 Level	-115.51 ± 51.69	-138.70 ± 109.50	-99.58 ± 10.82	-697.83 ± 46.95	-291.21 ± 80.64
InfoVis-Vast	1 Level	0.20 ± 0.00	-257.31 ± 0.00	-33.92 ± 0.00	-387.50 ± 38.85	-334.82 ± 0.00
	2 Level	0.08 ± 0.08	-310.04 ± 41.24	-59.86 ± 4.40	-403.23 ± 39.74	-425.13 ± 62.71
	3 Level	-4.91 ± 11.81	-432.98 ± 54.56	-62.42 ± 10.16	-430.37 ± 38.30	-481.76 ± 47.85
Pinterest	1 Level	-70.83 ± 18.60	-533.33 ± 0.00	-71.06 ± 0.00	-576.22 ± 30.36	-572.82 ± 0.00
	2 Level	-99.54 ± 42.53	-233.10 ± 75.69	-23.58 ± 18.07	-551.13 ± 30.74	-597.43 ± 29.30
	3 Level	-130.69 ± 50.72	-92.76 ± 74.58	-61.31 ± 11.70	-514.51 ± 37.95	-255.41 ± 56.51
TripAdvisor	1 Level	-19.15 ± 0.00	-457.78 ± 0.00	-74.75 ± 0.00	-583.20 ± 32.28	-493.97 ± 0.00
	2 Level	-33.33 ± 11.39	-224.06 ± 80.71	-33.34 ± 22.71	-590.83 ± 22.31	-651.94 ± 13.39
	3 Level	-58.03 ± 27.52	-113.90 ± 82.96	-72.45 ± 8.90	-581.31 ± 30.75	-132.31 ± 43.91
Tweets	1 Level	-80.04 ± 0.00	-826.50 ± 0.00	-94.64 ± 0.00	-683.25 ± 68.75	-832.73 ± 0.00
	2 Level	-68.88 ± 24.80	-251.61 ± 64.09	-79.45 ± 9.09	-673.96 ± 66.61	-805.36 ± 25.10
	3 Level	-98.40 ± 30.67	-106.99 ± 62.55	-93.81 ± 8.80	-656.36 ± 50.74	-260.31 ± 53.26
Uber	1 Level	-34.86 ± 0.00	-555.34 ± 0.00	-94.14 ± 0.00	-658.81 ± 52.17	-577.00 ± 0.00
	2 Level	-40.37 ± 10.30	-576.02 ± 92.22	-85.52 ± 17.19	-678.75 ± 57.32	-673.40 ± 21.48
	3 Level	-79.09 ± 40.06	-172.80 ± 117.45	-95.64 ± 7.48	-689.03 ± 53.47	-386.45 ± 102.42
Whatsapp	1 Level	-56.74 ± 0.00	-597.47 ± 0.00	-55.41 ± 0.00	-604.01 ± 14.66	-686.37 ± 0.00
	2 Level	-59.06 ± 15.53	-147.96 ± 84.02	-23.31 ± 14.63	-577.14 ± 27.50	-571.32 ± 90.15
	3 Level	-110.30 ± 37.06	-47.66 ± 43.13	-50.93 ± 13.31	-527.78 ± 40.16	-188.80 ± 38.45

Table 15: Coherence results by level of hierarchy compared with uHTM strategies.

Datasets	Level	CluNMF	hLDA	hPAM	HSOC	KHTM
20News	1 Level	0.9863 ± 0.0000	0.1577 ± 0.0000	0.2338 ± 0.0000	0.2786 ± 0.0870	0.1059 ± 0.0000
	2 Level	0.9386 ± 0.0311	0.3358 ± 0.1395	0.3014 ± 0.0861	0.3041 ± 0.1177	0.2082 ± 0.0739
	3 Level	0.9495 ± 0.0319	0.4735 ± 0.1443	0.2090 ± 0.0526	0.2799 ± 0.0298	0.4542 ± 0.1123
ACM	1 Level	0.9552 ± 0.0185	0.1701 ± 0.0000	0.2192 ± 0.0000	0.1979 ± 0.0516	0.1960 ± 0.0000
	2 Level	0.9552 ± 0.0185	0.3193 ± 0.1018	0.1943 ± 0.0238	0.1875 ± 0.0507	0.0771 ± 0.0540
	3 Level	0.9682 ± 0.0071	0.5861 ± 0.0911	0.1735 ± 0.0442	0.1874 ± 0.0473	0.4690 ± 0.0607
AngryBirds	1 Level	0.9729 ± 0.0000	0.0749 ± 0.0000	0.2516 ± 0.0000	0.1708 ± 0.0211	0.0530 ± 0.0000
	2 Level	0.9486 ± 0.0135	0.4849 ± 0.0657	0.5608 ± 0.0777	0.2006 ± 0.0232	0.2076 ± 0.1048
	3 Level	0.8887 ± 0.0504	0.5711 ± 0.0406	0.3415 ± 0.0783	0.2280 ± 0.0226	0.5055 ± 0.0346
Dropbox	1 Level	0.9819 ± 0.0000	0.0522 ± 0.0000	0.2795 ± 0.0000	0.1385 ± 0.0262	0.0565 ± 0.0000
	2 Level	0.9184 ± 0.0312	0.4441 ± 0.0768	0.3911 ± 0.0873	0.1589 ± 0.0263	0.2048 ± 0.0490
	3 Level	0.8980 ± 0.0459	0.6009 ± 0.0645	0.2389 ± 0.0753	0.1839 ± 0.0289	0.4131 ± 0.0378
Evernote	1 Level	0.9760 ± 0.0000	0.0522 ± 0.0000	0.1485 ± 0.0000	0.1105 ± 0.0225	0.0698 ± 0.0000
	2 Level	0.9659 ± 0.0092	0.4177 ± 0.0706	0.2694 ± 0.0918	0.1190 ± 0.0247	0.0398 ± 0.0116
	3 Level	0.9341 ± 0.0334	0.5780 ± 0.0704	0.1419 ± 0.0348	0.1268 ± 0.0213	0.4499 ± 0.0544
Facebook	1 Level	0.9330 ± 0.0141	0.0167 ± 0.0000	0.1863 ± 0.0000	0.1035 ± 0.0601	0.0448 ± 0.0000
	2 Level	0.9017 ± 0.0402	0.3331 ± 0.0908	0.2471 ± 0.1057	0.1100 ± 0.0463	-0.0237 ± 0.0082
	3 Level	0.8627 ± 0.0527	0.6086 ± 0.0530	0.1418 ± 0.0660	0.1166 ± 0.0406	0.4865 ± 0.0434
InfoVis-Vast	1 Level	1.0001 ± 0.0000	0.0379 ± 0.0000	0.0353 ± 0.0000	0.1610 ± 0.0516	0.0128 ± 0.0000
	2 Level	1.0000 ± 0.0001	0.0441 ± 0.0037	0.1949 ± 0.0486	0.1576 ± 0.0521	0.0666 ± 0.0430
	3 Level	0.9929 ± 0.0198	0.1938 ± 0.0475	0.1124 ± 0.0472	0.1666 ± 0.0489	0.1750 ± 0.0660
Pinterest	1 Level	0.9042 ± 0.0258	0.0161 ± 0.0000	0.2502 ± 0.0000	0.1391 ± 0.0376	0.0051 ± 0.0000
	2 Level	0.8634 ± 0.0488	0.4311 ± 0.0594	0.5187 ± 0.1105	0.1700 ± 0.0362	0.1688 ± 0.0343
	3 Level	0.8246 ± 0.0503	0.5762 ± 0.0440	0.2818 ± 0.0669	0.2065 ± 0.0298	0.4001 ± 0.0303
TripAdvisor	1 Level	0.9775 ± 0.0000	0.0701 ± 0.0000	0.1002 ± 0.0000	0.1147 ± 0.0159	0.0656 ± 0.0000
	2 Level	0.9561 ± 0.0109	0.4652 ± 0.0746	0.4475 ± 0.1408	0.1350 ± 0.0193	0.1549 ± 0.0250
	3 Level	0.9232 ± 0.0342	0.5920 ± 0.0585	0.2589 ± 0.0599	0.1623 ± 0.0288	0.5115 ± 0.0420
Tweets	1 Level	0.9036 ± 0.0000	0.0149 ± 0.0000	0.1987 ± 0.0000	0.1756 ± 0.0715	0.0180 ± 0.0000
	2 Level	0.9176 ± 0.0311	0.4764 ± 0.0442	0.2763 ± 0.0422	0.1850 ± 0.0566	0.1019 ± 0.0236
	3 Level	0.8902 ± 0.0310	0.6029 ± 0.0231	0.2068 ± 0.0407	0.2010 ± 0.0441	0.4801 ± 0.0254
Uber	1 Level	0.9600 ± 0.0000	0.0195 ± 0.0000	0.1076 ± 0.0000	0.0845 ± 0.0319	0.0232 ± 0.0000
	2 Level	0.9544 ± 0.0091	0.2844 ± 0.0916	0.2504 ± 0.0871	0.0953 ± 0.0313	0.0090 ± 0.0160
	3 Level	0.9069 ± 0.0419	0.5927 ± 0.0658	0.1297 ± 0.0408	0.1073 ± 0.0275	0.4246 ± 0.0657
Whatsapp	1 Level	0.9241 ± 0.0000	0.1105 ± 0.0000	0.3840 ± 0.0000	0.1574 ± 0.0222	0.0533 ± 0.0000
	2 Level	0.9208 ± 0.0201	0.5428 ± 0.0525	0.5285 ± 0.0741	0.1893 ± 0.0307	0.2394 ± 0.0770
	3 Level	0.8528 ± 0.0426	0.5951 ± 0.0166	0.3846 ± 0.0618	0.2214 ± 0.0324	0.5295 ± 0.0161

Table 16: NPMI results by level of hierarchy compared with uHTM strategies.

Datasets	Level	CluNMF	hLDA	hPAM	HSOC	KHTM
20News	1 Level	1.0232 ± 0.0000	0.9866 ± 0.0000	1.0314 ± 0.0000	0.3025 ± 0.2030	0.0346 ± 0.0000
	2 Level	1.1925 ± 0.1096	1.3775 ± 0.1562	1.1515 ± 0.0843	0.3221 ± 0.1925	0.2698 ± 0.1107
	3 Level	1.2060 ± 0.1183	1.4500 ± 0.1361	1.1308 ± 0.0857	0.3356 ± 0.2301	0.2137 ± 0.1775
ACM	1 Level	1.0955 ± 0.1047	0.9653 ± 0.0000	1.1710 ± 0.0000	0.6783 ± 0.2517	0.4138 ± 0.0000
	2 Level	1.0955 ± 0.1047	1.3447 ± 0.1152	1.1661 ± 0.0721	0.6292 ± 0.2627	0.6458 ± 0.2348
	3 Level	1.0320 ± 0.0716	1.4782 ± 0.0873	1.1255 ± 0.1006	0.6373 ± 0.2791	0.1470 ± 0.1382
AngryBirds	1 Level	0.9416 ± 0.0000	1.2087 ± 0.0000	1.2447 ± 0.0000	1.1833 ± 0.0379	1.2514 ± 0.0000
	2 Level	1.1148 ± 0.1406	1.2806 ± 0.0502	1.2878 ± 0.0569	1.1797 ± 0.0343	1.1949 ± 0.0341
	3 Level	1.1532 ± 0.1120	1.3300 ± 0.0229	1.2225 ± 0.0777	1.1822 ± 0.0410	1.2626 ± 0.0256
Dropbox	1 Level	0.9895 ± 0.0000	1.1067 ± 0.0000	1.2108 ± 0.0000	1.1641 ± 0.0264	1.1113 ± 0.0000
	2 Level	1.1191 ± 0.0966	1.2912 ± 0.0507	1.2193 ± 0.0613	1.1590 ± 0.0426	1.1794 ± 0.0472
	3 Level	1.1489 ± 0.0905	1.3459 ± 0.0318	1.1751 ± 0.0889	1.1747 ± 0.0432	1.3099 ± 0.0288
Evernote	1 Level	0.9707 ± 0.0000	1.1173 ± 0.0000	1.1561 ± 0.0000	1.1660 ± 0.0490	1.0959 ± 0.0000
	2 Level	1.0048 ± 0.0658	1.3185 ± 0.0671	1.1936 ± 0.0806	1.1725 ± 0.0385	1.1449 ± 0.0305
	3 Level	1.1109 ± 0.1249	1.3876 ± 0.0475	1.1397 ± 0.0606	1.1916 ± 0.0452	1.3306 ± 0.0464
Facebook	1 Level	1.0722 ± 0.1251	1.1125 ± 0.0000	1.2397 ± 0.0000	1.1611 ± 0.0444	1.1435 ± 0.0000
	2 Level	1.1471 ± 0.1405	1.2787 ± 0.0575	1.1766 ± 0.0346	1.1707 ± 0.0522	1.1124 ± 0.0168
	3 Level	1.1993 ± 0.1170	1.4197 ± 0.0314	1.1573 ± 0.0550	1.1836 ± 0.0594	1.3036 ± 0.0319
InfoVis-Vast	1 Level	1.0569 ± 0.0000	1.0817 ± 0.0000	1.0517 ± 0.0000	1.1783 ± 0.0507	1.1137 ± 0.0000
	2 Level	1.0604 ± 0.0589	1.0673 ± 0.0066	1.2105 ± 0.0526	1.1824 ± 0.0490	1.1466 ± 0.0570
	3 Level	1.1091 ± 0.0881	1.2228 ± 0.0440	1.1617 ± 0.0639	1.2000 ± 0.0504	1.1820 ± 0.0457
Pinterest	1 Level	1.1468 ± 0.1059	1.1385 ± 0.0000	1.1699 ± 0.0000	1.1522 ± 0.0531	1.1640 ± 0.0000
	2 Level	1.1850 ± 0.0991	1.2690 ± 0.0321	1.2962 ± 0.0771	1.1617 ± 0.0495	1.1918 ± 0.0331
	3 Level	1.2467 ± 0.0772	1.2938 ± 0.0236	1.2070 ± 0.0655	1.1892 ± 0.0441	1.2269 ± 0.0244
TripAdvisor	1 Level	0.9554 ± 0.0000	1.0859 ± 0.0000	1.1188 ± 0.0000	1.1275 ± 0.0247	1.1113 ± 0.0000
	2 Level	1.0465 ± 0.0663	1.2975 ± 0.0563	1.2573 ± 0.0611	1.1441 ± 0.0320	1.1783 ± 0.0327
	3 Level	1.1155 ± 0.1087	1.3782 ± 0.0345	1.1744 ± 0.0646	1.1534 ± 0.0310	1.3204 ± 0.0206
Tweets	1 Level	0.9102 ± 0.0000	1.0165 ± 0.0000	1.2447 ± 0.0000	1.2161 ± 0.0827	1.1532 ± 0.0000
	2 Level	1.0231 ± 0.0463	1.3655 ± 0.0525	1.2616 ± 0.0339	1.2282 ± 0.0679	1.1491 ± 0.0227
	3 Level	1.0592 ± 0.1151	1.4350 ± 0.0246	1.2092 ± 0.0661	1.2242 ± 0.0650	1.3304 ± 0.0323
Uber	1 Level	1.0136 ± 0.0000	1.1040 ± 0.0000	1.1916 ± 0.0000	1.1716 ± 0.0340	1.1249 ± 0.0000
	2 Level	1.0353 ± 0.0867	1.2679 ± 0.0557	1.2252 ± 0.0630	1.1684 ± 0.0371	1.1468 ± 0.0195
	3 Level	1.1431 ± 0.1329	1.3794 ± 0.0360	1.1277 ± 0.0600	1.1663 ± 0.0396	1.3048 ± 0.0474
Whatsapp	1 Level	0.9615 ± 0.0000	1.1816 ± 0.0000	1.2917 ± 0.0000	1.1685 ± 0.0421	1.1781 ± 0.0000
	2 Level	0.9975 ± 0.0661	1.2235 ± 0.0270	1.2748 ± 0.0427	1.1675 ± 0.0401	1.2057 ± 0.0394
	3 Level	1.1380 ± 0.1030	1.2257 ± 0.0110	1.2095 ± 0.0644	1.1773 ± 0.0430	1.2991 ± 0.0178

Table 17: W2V-L1 results by level of hierarchy compared with uHTM strategies.