Scalable Text Vectorization with Hyperdimensional Computing Through Selective Word Encoding

Timur Mudarisov
University of Luxembourg
Luxembourg
timur.mudarisov@uni.lu

Zsofia Kraussl
Bayes Business School
London
zsofia.kraussl@bayes.ac.uk

Evgeny Polyachenko
University of Luxembourg
Luxembourg
evgeny.polyachenko@uni.lu

Enriqueta Patricia Becerra Sanchez
University of Luxembourg
Luxembourg
enriqueta.becerra@uni.lu

Tatiana Petrova
University of Luxembourg
Luxembourg
tatiana.petrova@uni.lu

Radu State
University of Luxembourg
Luxembourg
radu.state@uni.lu

Abstract

Hyperdimensional Computing (HDC) is a promising approach for various machine learning tasks. In this work, we focus on its application to encoding large text datasets, where the *curse of dimensionality* presents a significant challenge. To mitigate this issue, we employ compression techniques that are based on classical models such as Term Frequency-Inverse Document Frequency (TF-IDF) and Latent Dirichlet Allocation (LDA). We derive theoretical expressions for Compression Rate, Jensen-Shannon Divergence, and ROUGE score, which quantify text size reduction, preservation of word distributions, and retention of key information, respectively. These expressions are validated using the IMDB, arXiv, and AG News datasets. Our results demonstrate that TF-IDF compression can reduce the encoded text size to 10% (or less in some cases) of the original input while also achieving slightly worse distinguishability between classes in classification tasks.

1 Introduction

Hyperdimensional Computing (HDC) is a machine learning approach inspired by principles of neural computation. It represents and manipulates data through high-dimensional vectors, typically in the order of thousands or millions, enabling information processing and storage. This methodology exhibits inherent robustness to noise, offers efficient learning capabilities, and effectively handles complex, unstructured data (Kanerva, 2009). HDC has gained considerable interest in emerging applications, such as robotics and health diagnostics, alongside established areas including data center recommendation systems (Mitrokhin et al., 2019; Neubert and Schubert, 2021; Yunhui et al., 2021). This increasing adoption and interest highlights the need for a robust theoretical justification. To address this, researchers have investigated HDC from different perspectives. These studies include an in-depth examination of its geometric characteristics (Pourmand et al., 2024), a comprehensive analysis of its algebraic foundations (Yu et al., 2024), and a detailed investigation of encoding structures used within HDC systems (Thomas et al., 2021). Each perspective contributes to a deeper understanding of HDC and its potential applications.

Kanerva (2009) identified several valuable aspects of different HDC realizations. These include their robustness to noise, which allows HDC to maintain performance despite disruptions. Their inherent transparency also helps the understanding and interpretation of results. Furthermore, HDC exhibit useful distributed properties, which enable efficient parallel processing, for example using GPUs. HDC have been successfully applied in various scientific fields (Rahimi et al., 2019; Kanerva, 2009), and their application to Natural Language Processing (NLP) tasks is of particular interest. Specifically, Kleyko et al. (2023) demonstrated successful applications of HDC to translation, sentence similarity, and topic classification problems. However, Thomas et al. (2021) pointed out important limitations of basic HDC. Among these, a critical challenge is the *curse of dimensionality*. This effect describes how increases in data size can cause an exponential rise in vector space dimensionality, complicating analysis and processing.

To address the challenge of the curse of dimen-

sionality in HDC, we propose using text compression techniques. In this paper, we aim at exploring two classical techniques for text compression: TF-IDF selection (Spärck Jones, 1972) and LDA (Blei et al., 2003). Our contribution to the state-of-the-art in HDC is threefold: First, in Section 2 we introduce a novel model – *compression HDC* (CHDC) which combines a theory-based encoding procedure with data compression using TF-IDF or LDA. This model allows encoding information efficiently while reducing the size of representations. Second, we analyze the compression effect of these techniques (Section 3.1), providing theorems that estimate the compression rate. Third, we examine the encoding effect of the binary uniform HDC (Section 3.2), showing that our results are robust to different conditions. In Section 4, we experimentally validate our theoretical findings, for the quality of the proposed compression and encoding processes. Finally, Section 5 wraps up and discusses prospects.

2 Model Setup

The scheme of our proposed model is presented in Figure 1. Before any text analysis is performed, a standard procedure of pre-processing is used and is therefore not shown in the scheme. This procedure involves four steps applied to a large text (document): first, only letters and numbers are retained; second, the text is broken down into words; third, lemmatization is applied, which reduces words to their base or dictionary form (lemma); and finally, stemming is applied, which reduces words to their root form.

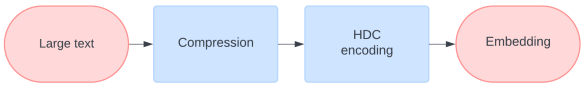

Figure 1: Workflow of the compression HDC model, illustrating the processing of a large text using text compression and HDC encoding (blue), to produce a final embedding.

The core of our proposed compression HDC model is defined by two components: compression and HDC encoding. These components are detailed in Sections 2.1 and 2.2, respectively.

2.1 Compression procedure

Let $\mathcal{W} = \{w_1, \ldots, w_M\}$ represent a set of M unique words and corpus $\mathcal{D} = \{d_1, \ldots, d_N\}$ is a set of N documents. Given these sets $(\mathcal{W}, \mathcal{D})$, our goal is to reduce the number of words in each document by focusing on the most informative ones. To achieve this, we assign a score to each word and extract the set of word-score pairs $\{(w, s_w)\}$. For the TF-IDF-based compression, we define the score as follows:

Definition 1. *The TF-IDF score for a word w_i in a corpus $\mathcal{D}$ is defined as:*

$$s_w = \mathrm{ts}(w, \mathcal{D}) = \frac{1}{N} \sum_{j=1}^{N} f_{w,j} \ln \frac{N}{N_w} , \tag{1}$$

where $f_{w,j}$ is the frequency of word w in document d_j, N_w is the number of documents in $\mathcal{D}$ containing word w.

Note that our definition differs from the standard TF-IDF definition, which depends on w, d and $\mathcal{D}$ and does not contain averaging over documents.

Latent Dirichlet Allocation (LDA) assumes that documents are represented as bags of words, where each document is a mixture of T topics, with T being a predefined number of topics. The probability of a word w belonging to topic t is denoted as $\phi_{t,w}$. The matrix $\Phi = \{\phi_1, \ldots, \phi_T\} \in \mathbb{R}^{T \times M}$, where each ϕ_t represents the probability distribution of words for topic t, is determined by maximizing the likelihood function $\mathbb{P}(\mathcal{W}, \mathcal{D} | \Phi, \alpha)$, and $\alpha \in \mathbb{R}_+^T$ are the parameters of the Dirichlet distribution (Blei et al., 2003). Based on the LDA model, we define the score as follows:

Definition 2. *The LDA-based score for a word w in topic t is defined as:*

$$s_{t,w} = \phi_{t,w} . \tag{2}$$

We consider the documents unordered and refer to them interchangeably using either the index j or the document d itself, as an element of $\mathcal{D}$. For words and word-related quantities, we will refer to them interchangeably using either the word w itself or the index i, specifying the ordering when necessary. Thus, for example, $f_{i,j}$ and $f_{w,d}$ denote the same quantity.

We present the following *compression criteria*. For TF-IDF-based compression, we select the p-quantile of words with the highest scores from the set $\{(w, s)\}_{w \in \mathcal{W}}$, resulting in a reduced dictionary $\mathcal{W}_p$ containing approximately pM words. For LDA-based compression, we select the top pM words from each topic, based on their topic probabilities $s_{w,t}$. Because each word has a probability of belonging to every topic, the resulting

reduced dictionary $\mathcal{W}_p$ typically contains fewer than TpM words. Subsequently, we create a new set of compressed documents $\mathcal{D}' = \{d'_1, \ldots, d'_N\}$, where each d'_j is formed by combining words from $\mathcal{W}_p$, preserving the most important words of the original document and their sequential order within each document.

To evaluate the compression quality, we introduce three classical performance metrics:

1. **Compression rate**. A standard metric in compression theory, defined as the ratio:

$$\mathrm{CR} = \frac{\sum_{j=1}^{N} |d'_j|}{\sum_{j=1}^{N} |d_j|}\,, \tag{3}$$

where $|d_j|$ and $|d'_j|$ denote the total number of non-unique words in the uncompressed document d_j and the compressed document d'_j, respectively. This metric directly quantifies the reduction in text size achieved by compression.

2. **Jensen-Shannon divergence.** For distributions p and q, the Jensen-Shannon divergence (JSD) measures the dissimilarity between word distributions and is defined as:

$$\mathrm{JSD}(p||q) = \frac{1}{2}\left[D_{\mathrm{KL}}(p||m) + D_{\mathrm{KL}}(q||m)\right]\,, \tag{4}$$

where D_{KL} is the Kulback-Leibler divergence, $m = (p+q)/2$. For TF-IDF compression, we calculate the JSD between the average word frequencies in the original and compressed documents, defined as:

$$p_i = \frac{1}{N}\sum_{j=1}^{N} f_{i,j}\,, \quad q_i = \frac{1}{N}\sum_{j=1}^{N} f'_{i,j}\,, \tag{5}$$

where $f_{i,j}$ and $f'_{i,j}$ are the frequencies of word w_i in documents d_j and d'_j, respectively.

For LDA compression, we use the average JSD across all topics, defined as:

$$\overline{\mathrm{JSD}}(p||q) = \sum_{t=1}^{T} \pi_t\, \mathrm{JSD}(p_t, q_t)\,, \tag{6}$$

$$\pi_t = \frac{1}{N}\sum_{j=1}^{N} z_{t,d_j}\,, \tag{7}$$

where $z_{t,d}$ is an indicator function that equals 1 if topic t is the most probable topic for document d, and zero otherwise. The densities p_t and q_t are defined as:

$$p_{t,i} = \frac{1}{N_t}\sum_{j=1}^{N} f_{i,j}\, z_{t,d_j}\,, \tag{8}$$

$$q_{t,i} = \frac{1}{N_t}\sum_{j=1}^{N} f'_{i,j}\, z_{t,d_j}\,, \tag{9}$$

with $f_{i,j}$ and $f'_{i,j}$ given in (5), and N_t is the number of documents for which topic t is the most probable one. Further details on the properties of JSD are available in Lin (1991). This metric allows us to evaluate how well the compressed documents retain the original word distributions.

3. **ROUGE score.** As a summarization metric, used to evaluate the quality of text summarization, we use the ROUGE-LCS score, introduced in Lin (2004), where $\mathrm{LCS}(r, s)$ denotes the length of the longest sequence of words that appear in both r and s in the same order. The ROUGE-F1 score is defined as:

$$\text{ROUGE-F1} = 2\,\frac{R\,P}{R+P}\,, \tag{10}$$

where recall $R = |\,\mathrm{LCS}(r,s)|/|r|$ and precision $P = |\,\mathrm{LCS}(r,s)|/|s|$; $|r|$, $|s|$, and $|\,\mathrm{LCS}(r,s)|$ are the word counts in the corresponding sequences. This metric is used to assess how well the compressed documents retain the key information of the original documents.

2.2 Encoding procedure

We now describe the steps of the encoding procedure, following the work by Kanerva (1988):

1. We consider the English alphabet plus digits, denoted as $\mathcal{A}$, and assign to each element $a_k \in \mathcal{A}$ a random vector $\phi(a_k)$ from the space $\mathcal{H} = \{\pm 1\}^D$, where D is the dimension of the space. In this vector space, we define a coordinate-wise multiplication operation $\otimes$ and a coordinate-wise sign operation $\oplus$. The multiplication is a simple coordinate-wise product, while the sign operation is applied after a coordinate-wise summation, with the sign of zero defined as 1.

2. We use word-wise encoding. To encode a word, we apply a permutation operation ρ to each character's vector $\phi(a_k)$, shifting all but the first coordinate to the left. The encoding vector for word w_i is then:

$$\phi(w_i) = \bigotimes_{0 \le k < |w_i|} \rho^k(\phi(a_k))\,, \tag{11}$$

where $|w_i|$ is the number of characters in word w_i.

3. The document encoding is obtained by applying the sign operation to the coordinate-wise summation of all word vectors:

$$\phi(d) = \bigoplus_{i=1}^{|d|} \phi(w_i)\,. \tag{12}$$

The outcome of this encoding procedure is a function $\phi(d)$ that maps a text to the vector space $\mathcal{H}$.

3 Theoretical analysis

We divide our theoretical analysis into two main components: *compression* and *encoding*, based on the compression HDC model (Figure 1) and the previous section. These components are supported by intuition, assumptions and theorems in the following subsections.

3.1 Compression

In this section, we present our compression analysis separately for TF-IDF and LDA-based approaches. The original TF-IDF and LDA statistics were introduced by (Aizawa, 2003) and (Blei et al., 2003), respectively.

3.1.1 TF-IDF part

We analyze the TF-IDF score $\mathrm{ts}(w_i, \mathcal{D})$ as a random variable. The randomness stems from the frequency $f_{i,j}$ and the number of documents N_{w_i} containing the word w_i. The frequency $f_{i,j}$ is related to the number of occurrences $n_{i,j}$ of word w_i in document d_j as $n_{i,j} = f_{i,j}|d_j|$. We can represent the documents schematically as:

$$d_j = \underbrace{w_1 \dots w_1}_{n_{1,j}} \dots \underbrace{w_M \dots w_M}_{n_{M,j}}\,. \tag{13}$$

Thus, each document can be considered as a random vector $(n_{1,j}, n_{2,j}, ..., n_{M,j})$. To proceed with our analysis, we make the following assumptions:

Assumption 1 (Poisson-like distribution and independence across documents). *To model the TF-IDF distribution, we assume that the number of occurrences $n_{i,j}$ of word w_i in document d_j are independent of the document d_j and follows a distribution $Dist(\lambda_i)$, where:*

$$\mathbb{P}(n_{i,j} = k) = \begin{cases} 1 - f(\lambda_i)\,, & k = 0\,; \\ f(\lambda_i)\dfrac{\lambda_i^k e^{-\lambda_i}}{k!(1-e^{-\lambda_i})}\,, & k > 0\,. \end{cases} \tag{14}$$

Here, $f(\lambda_i)$ is an auxiliary function introduced to make our theoretical analysis tractable and ensure a monotonically growing TF-IDF approximate estimate, prioritizing words with larger λ_i for encoding.

The next assumption allows us to exclude randomness from the TF part:

Assumption 2 (Average frequency). *The TF part can be fixed at p_i, by approximating the average frequency as:*

$$\frac{1}{N}\sum_{j=1}^{N} f_{i,j} = \frac{1}{N}\sum_{j=1}^{N}\frac{n_{i,j}}{|d_j|} \approx \frac{\mathbb{E}n_i}{\mathbb{E}|d_j|} = p_i\,, \tag{15}$$

where $\mathbb{E}|d_j| = \sum_{i=1}^{M}\lambda_i$.

Thus randomness retains only in the IDF part, i.e. in N_w. To estimate the number of documents where word w occurs at least once, we have:

$$N_w = \sum_{j=1}^{N} \mathbf{1}(w \in d_j)\,, \tag{16}$$

which is a sum of N i.i.d. Bernoulli variables $\mathrm{Bern}(q_w)$ with $q_w = 1 - \exp(-\lambda_w)$. Hence, the expectation of N_w is $q_w N$, and for the TF-IDF approximate we obtain:

$$\widetilde{\mathrm{ts}}(w) = -\frac{\lambda_w f(\lambda_w)}{(1-e^{-\lambda_w})\mathbb{E}|d|}\ln(1-e^{-\lambda_w})\,. \tag{17}$$

To ensure a monotonically growing TF-IDF approximation, we make the next assumption:

Assumption 3 (Function $f(x)$). *Function $f(x)$ is defined as:*

$$f(\lambda) = \frac{\lambda}{1+\lambda}(1-e^{-\lambda})\,. \tag{18}$$

This results in the following score approximate expectation:

$$\widetilde{\mathrm{ts}}(w) = -\frac{\lambda_w^2}{(1+\lambda_w)\mathbb{E}|d|}\ln\left[\frac{\lambda_w}{1+\lambda_w}(1-e^{-\lambda_w})\right] \tag{19}$$

with the asymptotic behavior $\widetilde{\mathrm{ts}}(w)\mathbb{E}|d| = 1 - 3/(2\lambda) + \mathcal{O}(\lambda^{-2})$, i.e. attaining gradually 1 from below.

Figure 2 illustrates the true TF-IDF score (1) for IMDB dataset and our approximate expectation $\widetilde{\mathrm{ts}}(w)$ as a function of the parameter estimate $\widehat{\lambda}_w$, obtained using the method of moments from the equation:

$$n_w \equiv \frac{1}{N}\sum_{j=1}^{N} n_{w,j} = \frac{\widehat{\lambda}_w f(\widehat{\lambda}_w)}{1-e^{-\widehat{\lambda}_w}} \tag{20}$$

(here and below, estimators of random variables are denoted with a wide hat). As can be observed, $\widetilde{\mathrm{ts}}(w)$

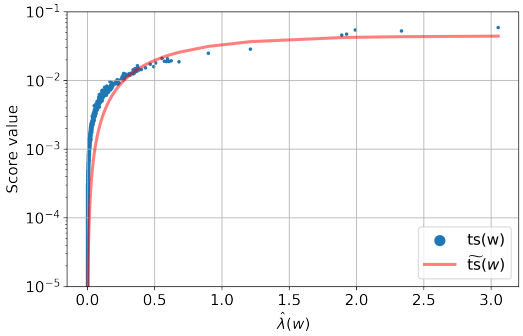

Figure 2: Comparison of the true TF-IDF statistics $\text{ts}(w)$ (1) for IMDB dataset and its approximate expectation $\widetilde{\text{ts}}(w)$ (19).

grows monotonically, as does the average true TF-IDF. However, the true TF-IDF values exhibit a noticeable vertical scatter (see blue points at $\lambda_w < 0.4$) due to the inherent randomness of the true TF-IDF score.

The compression method outlined in Section 2.1 selects words with the largest TF-IDF score:

$$\mathcal{W}_p = \{w \in \mathcal{W} : \text{ts}(w) \geq \text{ts}_{(\lceil (1-p)M \rceil)}\}. \tag{21}$$

Here and below $X_{(k)}$ denote is the k-th order statistic of $\{X(w_1), \ldots, X(w_M)\}$. Due to the complexity of $\text{ts}(w)$, we use expectation $\widetilde{\text{ts}}(w)$ to select the pM words with the highest values of $\widehat{\lambda}_w$:

$$\widehat{\mathcal{W}}_p = \{w_i \in \mathcal{W} : \widehat{\lambda}_i \geq \widehat{\lambda}_*\}, \tag{22}$$

where $\widehat{\lambda}_* \equiv \widehat{\lambda}_{(\lceil (1-p)M \rceil)}$ is the minimal value $\widehat{\lambda}_w$ of the word w included in set $\widehat{\mathcal{W}}_p$. Although $\mathcal{W}_p$ and $\widehat{\mathcal{W}}_p$ are not identical due to the randomness of $\text{ts}(w_i)$ and $\widehat{\lambda}_w$, the monotonicity of $\widetilde{\text{ts}}(w)$ implies that both sets will contain the same words, except for those in the vicinity of $\widehat{\lambda}_*$, where some words will be randomly added and others excluded from $\mathcal{W}_p$. To simplify our analysis, we assume that the sets $\mathcal{W}_p$ and $\widehat{\mathcal{W}}_p$ differ negligibly:

Assumption 4 (Negligible difference in selected words). *We assume that $\mathcal{W}_p$ and $\widehat{\mathcal{W}}_p$ differ negligibly.*

For the theorems, we require an informational inequality (proof follows from Pinsker's inequality and Lin, 1991):

Lemma 1. *For Jensen-Shannon divergence, we have:*

$$\frac{1}{4}\left[V^2(p,m) + V^2(q,m)\right] \leq \text{JSD}(p||q) \leq \frac{1}{2}V(p,q), \tag{23}$$

where $V(p,q) = \sum_i |p_i - q_i|$ and $m = (p+q)/2$.

We now formulate the theorems (see Appendix A.1.1 for the proof).

Theorem 1 (TF-IDF compression)**.** *Based on assumptions 1–4, we have the consistent estimators for* CR, JSD$(p||q)$ *and* ROUGE-F1:

$$\widehat{\text{CR}} = \frac{\sum_{w \in \widehat{\mathcal{W}}_p} g(\widehat{\lambda}_w)}{\sum_{w \in \mathcal{W}} g(\widehat{\lambda}_w)}, \tag{24}$$

$$\widehat{\text{JSD}}(p||q) = \frac{1}{2}\left[\sum_{w \in \widehat{\mathcal{W}}_p} \widehat{p}_w \ln\left(\frac{2\widehat{\text{CR}}}{\widehat{\text{CR}}+1}\right)\right] + \frac{\ln 2}{2}\sum_{w \in \mathcal{W}/\widehat{\mathcal{W}}_p} \widehat{p}_w + \frac{1}{2}\left[\sum_{w \in \widehat{\mathcal{W}}_p} \frac{\widehat{p}_w}{\widehat{\text{CR}}} \ln\left(\frac{2}{1+\widehat{\text{CR}}}\right)\right], \tag{25}$$

$$\widehat{\text{ROUGE-F1}} = 2\frac{\widehat{\text{CR}}}{\widehat{\text{CR}}+1}, \tag{26}$$

where $g(x) = x^2/(1+x)$ and $\widehat{p}_w = g(\widehat{\lambda}_w)/\sum_{w \in \mathcal{W}} g(\widehat{\lambda}_w)$.

Theorem 2 (Quantile criteria)**.** *Under assumptions 1–4, the TF-IDF compression model with p-quantile criteria has the following bounds from Table 1.*

3.1.2 LDA part

We now examine the LDA compression procedure. For a fixed topic t, the distribution of words is a Dirichlet random variable, $\Phi_t \sim Dir(\alpha)$, where α is a vector of parameters $(\alpha_1, \ldots, \alpha_M)$ (see Blei et al., 2003, for details). As outlined in Section 2.1, we define the set:

$$\mathcal{W}_{t,p} = \{w_i \in \mathcal{W} : \Phi_{t,w} \geq \Phi_{t,(\lceil (1-p)M \rceil)}\}, \tag{27}$$

where $\Phi_{t,w}$ is the probability of word w belonging to topic t. To determine the distribution of $\Phi_{t,(\lceil (1-p)M \rceil)}$, we need the marginal distributions of Φ_{t,w_i}.

Lemma 2. *If $\Phi = (\Phi_1, \ldots, \Phi_M) \sim Dir(\alpha)$, then its marginal distributions are beta distributed random variables:*

$$\Phi_i \sim Beta\left(\alpha_i, \sum_{k=1}^{M} \alpha_k - \alpha_i\right). \tag{28}$$

This lemma allows us to identify and generalize the object of our interest. Applying the same conceptual approach as in the TF-IDF part, we focus on the quantile value of the $(\Phi_{t,1}, \ldots, \Phi_{t,M})$, where each $\Phi_{t,i}$ is distributed as in (28).

The model has an additional parameter α, which we set to $(0.5, \ldots, 0.5)$, implying that we are unsure about word significance in topic t:

Assumption 5 (Non-significance). $\alpha = (0.5, \ldots, 0.5)$.

Under Assumption 5, we have a set of $Beta(0.5, 0.5[M-1])$ random variables. Using the same expectation approach as in the TF-IDF case, we focus on estimating $\mathbb{E}\Phi_{t,(k)}$. To proceed, we use the following lemma (see Arnold and Groeneveld, 1979, for the proof):

Lemma 3. *For i.i.d. random variables $X_1, \ldots, X_n$ with mean μ and variance σ^2, we have the following inequality:*

$$-\sigma\sqrt{\frac{n-k}{k}} \leq \mathbb{E}X_{(k)} - \mu \leq \sigma\sqrt{\frac{k-1}{n-k+1}}. \quad (29)$$

For $X \sim Beta(\alpha, \beta)$, we have:

$$\mu = \frac{\alpha}{\alpha+\beta} = M^{-1}, \quad (30)$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \quad (31)$$

$$\frac{M-1}{M^2(0.5M+1)} \approx 2M^{-2}. \quad (32)$$

Hence, we can estimate the bounds of $\mathbb{E}\Phi_{t,(\lceil(1-p)M\rceil)}$.

Before proceeding with the theorems, we clarify the distribution of the number of occurrences. Unlike the TF-IDF model, where we calculated $n_{i,j}$ directly, in the LDA model, we operate with $\Phi_{t,i}$ values. Therefore, we assume:

Assumption 6 (Poisson distribution). *For each topic t, we assume that the number of occurrences of each word w_i in a document d_t are independent random variables following the Poisson distribution:*

$$d_t = \underbrace{w_1 \ldots w_1}_{\upsilon_{t,1} \sim Pois(\Phi_{t,1}C)} \quad \ldots \quad \underbrace{w_M \ldots w_M}_{\upsilon_{t,M} \sim Pois(\Phi_{t,M}C)}, \quad (33)$$

where $\upsilon_{t,i}$ is the number of occurrences of word w_i in a document d_t belonging to topic t.

This assumption is quite strict, as it assumes a constant C that regulates the number of occurrences of each word in the document, and that this constant is the same for all topics. As we argue below, we use it to estimate the number of words in a document on a given topic.

Given a matrix of words in topic probabilities $\widehat{\Phi}_{t,w}$, we formulate the following theorems:

Theorem 3 (LDA compression estimators). *Under assumptions 5–6, we have asymptotically-unbiased estimators for* CR, JSD$(p||q)$, *and* ROUGE-F1*:*

$$\widehat{\mathrm{CR}} = \frac{\sum_{t=1}^{T}\pi_t \sum_{w\in\mathcal{W}_{p,t}} \widehat{\Phi}_{t,w}}{\sum_{t=1}^{T}\pi_t \sum_{w\in\mathcal{W}} \widehat{\Phi}_{t,w}}, \quad (34)$$

$$\widehat{\mathrm{JSD}}(p||q) = \frac{1}{2}\sum_{t=1}^{T}\pi_t\left[\sum_{w\in\mathcal{W}} \widehat{\Phi}_{t,w} \ln\left(\frac{2\widehat{\mathrm{CR}}}{\widehat{\mathrm{CR}}+1}\right)\right] + \frac{1}{2}\sum_{t=1}^{T}\pi_t\left[\sum_{w\in\mathcal{W}} \frac{\widehat{\Phi}_{t,w}}{\widehat{\mathrm{CR}}} \ln\left(\frac{2}{1+\widehat{\mathrm{CR}}}\right)\right] + \frac{\ln 2}{2}\sum_{t=1}^{T}\pi_t \sum_{w\in\mathcal{W}\setminus\mathcal{W}_p} \widehat{\Phi}_{t,w}, \quad (35)$$

$$\widehat{\mathrm{ROUGE\text{-}F1}} = 2\frac{\widehat{\mathrm{CR}}}{\widehat{\mathrm{CR}}+1} \quad (36)$$

with π_t defined by Eq. (7).

Theorem 4 (LDA compression bounds). *Under assumptions 5–6, the LDA compression model with p-quantile criteria has the following bounds from Table 1.*

3.2 Encoding

To prove the applicability of our proposed CHDC approach, we now turn to encoding implications and focus on estimating the quality of document analysis based on an average document size. As in the previous section, we consider documents as a bag of words (13). Consider now two documents, d_1 and d_2. Given the binary HDC encoding, we map our documents to the $\phi(d_1)$ and $\phi(d_2)$, according to the rules from Section 2.2. As pointed in (Kanerva, 1988), the HDC model should distinguish the vectors $\phi(d_1)$ and $\phi(d_2)$, which means that:

$$\langle\phi(d_1), \phi(d_2)\rangle \to 0 \quad (37)$$

with $D \to \infty$ (here $\langle ., . \rangle$ denotes the standard Euclidean dot-product). To estimate the effect of the encoding under fixed D, we propose to consider:

$$\mathbb{P}(\langle\phi(d_1), \phi(d_2)\rangle \geq \varepsilon D), \quad (38)$$

where D is the vector-space dimension, ε is small parameter that characterize distinguishability, ϕ is the encoding function, mentioned before. Notice that the $\phi(d)$ is a random vector, since we use a random binary HDC encoding. Therefore, we need to be sure that the probability of $\mathbb{P}(\langle\phi(d_1), \phi(d_2)\rangle > \varepsilon D)$ is low.

Let's rewrite the dot-product as follows:

$$\langle\phi(d_1), \phi(d_2)\rangle = \sum_{i=1}^{D} \phi_{1,i}\phi_{2,i} = \sum_{i=1}^{D} X_i, \quad (39)$$

Th.	CR	JSD	ROUGE-F1
Th. 2	$\left[pM\frac{\min_{w\in\mathcal{W}_p} g(\hat{\lambda}_w)}{\sum_{w\in\mathcal{W}} g(\hat{\lambda}_w)}; pM\frac{\max_{w\in\mathcal{W}_p} g(\hat{\lambda}_w)}{\sum_{w\in\mathcal{W}} g(\hat{\lambda}_w)}\right]$	$\left[\frac{1}{4}\left[\widehat{V}^2_{pm}+\widehat{V}^2_{qm}\right]; \frac{1}{2}\widehat{V}_{pq}\right]$	$\left[2\frac{\mathrm{CR}_{\min}}{1+\mathrm{CR}_{\min}}; 2\frac{\mathrm{CR}_{\max}}{1+\mathrm{CR}_{\max}}\right]$
Th. 4	$\left[p-\sqrt{\frac{2p}{1-p}}; p+p\sqrt{2(M-1)}\right]$	$\left[\frac{1}{4}\sum_t \widehat{\pi}_t\left[\widehat{V}^2_{t,pm}+\widehat{V}^2_{t,qm}\right]; \frac{1}{2}\sum_t \widehat{\pi}_t\widehat{V}_{t,pq}\right]$	$\left[2\frac{\mathrm{CR}_{\min}}{1+\mathrm{CR}_{\min}}; 2\frac{\mathrm{CR}_{\max}}{1+\mathrm{CR}_{\max}}\right]$

Table 1: Bounds for the performance metrics: compression rate (CR), Jensen-Shannon divergence (JSD), and ROUGE-F1 score, under TF-IDF (Theorem 2) and LDA (Theorem 4) compression.

where X_i are dependent Bernoulli-type random variables taking values in $\{\pm 1\}$, with $\gamma_i(X_1,\ldots,X_{i-1},\ X_{i+1},\ldots,X_D) = \mathbb{P}(X_i = 1|\{X_1,\ldots,X_D\}\setminus X_i)$. Unfortunately, we can't directly apply known techniques due to the possible dependency of the $\{X_i\}_{i=1}^D$. However, we propose the following lemma to overcome this problem (for proof, see Appendix A.2):

Lemma 4. *Assume $\{X_i\}_i$ are dependent random variables with Bernoulli-type distribution and $\mathbb{P}(X_i = 1|X_{i_1},\ldots,X_{i_k}) \le p$. Then there are $\{Y_i\}_i$ independent Bernoulli variables with $\mathbb{P}(Y_i = 1) = p$ and we have:*

$$\mathbb{P}\left(\sum_{i=1}^D X_i \ge \varepsilon D\right) \le \mathbb{P}\left(\sum_{i=1}^D Y_i \ge \varepsilon D\right). \quad (40)$$

The given lemma allows us to consider X_i as independent random variables with the same bound γ on its probability. To estimate the value of probability in (38), we propose using the following lemma (see Chernoff, 1952, for proof):

Lemma 5 (Chernoff bound)**.** *For a sum of independent random variables $X = \sum_i X_i$, we have:*

$$\mathbb{P}(X \ge a) \le \inf_{t>0}\left[e^{-ta}\prod_i \mathbb{E}e^{tX_i}\right]. \quad (41)$$

To justify the model, we formulate the following theorem (for the proof, see Appendix A.2):

Theorem 5. *The probability (38) is upper bounded by:*

$$\mathbb{P}(\langle\phi(d_1),\phi(d_2)\rangle \ge \varepsilon D) \le F(D,\gamma,\varepsilon), \quad (42)$$

where:

1. *The upper boundary:*

$$\ln F(D,\varepsilon,\gamma) = \frac{D}{2}(1-\varepsilon)\ln\left[\frac{1-\gamma}{\gamma}\frac{1+\varepsilon}{1-\varepsilon}\right] - D\ln\left[\frac{1+\varepsilon}{2\gamma}\right]. \quad (43)$$

2. *The Bernoulli probability γ satisfies the inequality:*

$$\frac{1}{2} < \gamma \le \frac{1}{2} + \binom{|d|}{\lceil |d|/2\rceil}\frac{1}{2^{|d|}} \approx \frac{1}{2} + \sqrt{\frac{2}{\pi|d|}}, \quad (44)$$

where $|d| = \mathbb{E}|d_i|$ is the average length of the document, the round brackets denote the binomial coefficient, and the asymptotical expansion in the r.h.s is obtained using Stirling's approximation.

The function F attains a maximum value of 1 when $\varepsilon + 1 = 2\gamma$. As we move away from this line, the function rapidly declines, with the decline becoming sharper as D increases. This implies that

$$\varepsilon \lesssim 2\sqrt{\frac{2}{\pi|d|}}. \quad (45)$$

For example, in the IMDB dataset, compression for $p = 0.1$ from an average document length of 122 words to 100 words increases ε by a factor of approximately $\sqrt{122/100} \approx 1.1$, just slightly worsening distinguishability.

4 Experiments

To verify our theoretical results, we propose a two-stage experimental setup, focusing on compression effect estimation and encoding results.

4.1 Compression analysis

We explore TF-IDF and LDA text compression techniques using Algorithm 1 (see A.4) applying it to IMDB reviews (Maas et al., 2011), AG News Dataset (Zhang et al., 2015), and arXiv dataset (Clement et al., 2019). Figure 3 (see A.3) presents the results, comparing direct calculations of the three metrics (CR, JSD and ROUGE-F1) with their theoretical expectations for different quantile parameters p. The green bounds show the possible ranges of metric scatter due to the randomness of word distributions (Theorems 2 and 4). The three upper panel rows demonstrate that TF-IDF

	TF-IDF			LDA		
D	$\widehat{\varepsilon}_{p=0.01}$	$\widehat{\varepsilon}_{p=0.1}$	$\widehat{\varepsilon}_{p=1}$	$\widehat{\varepsilon}_{p=0.01}$	$\widehat{\varepsilon}_{p=0.1}$	$\widehat{\varepsilon}_{p=1}$
256	0.17 ± 0.02	0.13 ± 0.01	0.12 ± 0.01	0.16 ± 0.02	0.13 ± 0.01	0.12 ± 0.01
1024	0.17 ± 0.02	0.13 ± 0.01	0.12 ± 0.01	0.16 ± 0.01	0.12 ± 0.01	0.12 ± 0.01
4096	0.17 ± 0.02	0.13 ± 0.01	0.12 ± 0.01	0.16 ± 0.01	0.12 ± 0.01	0.12 ± 0.01
16384	0.17 ± 0.02	0.12 ± 0.01	0.11 ± 0.01	0.16 ± 0.01	0.12 ± 0.01	0.11 ± 0.01

Table 2: Encoding analysis for TF-IDF and LDA compression techniques using the IMDB dataset. The table shows average scalar product values for dictionary compression parameters $p = 0.01$, 0.1, and 1 ($|d| \approx 60$, 100, 122, respectively) and vector space dimension D.

compression accurately captures all three metrics across all datasets and different values of p, because the relevant variables are directly observed and the assumptions are reasonable. In contrast, the three lower panels show that the LDA compression estimators perform worse, likely because the underlying distributional assumptions do not fully correspond to the actual distributions.

4.2 Encoding analysis

To validate the results in Section 3.2, we analyze how the encoding procedure impacts the distinguishability of randomly selected documents using the IMDB dataset. This dataset, which comprises two classes, simplifies our analysis (Algorithm 2) while still revealing key insights. We use Monte Carlo simulations with 100 iterations for the alphabet $\mathcal{A}$ and 100 iterations for document sampling (pairs from different classes), resulting in 10000 total iterations. Table 2 presents estimates of the parameter ε, defined as:

$$\widehat{\varepsilon}_p = D^{-1}\,\mathbb{E}|\langle \phi(d_{1,p}), \phi(d_{2,p})\rangle| \tag{46}$$

where $d_{1,p}$ and $d_{2,p}$ are randomly selected documents from different classes after compression, and p is the compression parameter. The table shows results for $p = 1$ (no compression, $|d| \approx 122$ words), $p = 0.1$ (medium compression, $|d| \approx 100$ words), and $p = 0.01$ (high compression, $|d| \approx 60$ words).

The estimates $\widehat{\varepsilon}_p$ are similar for TF-IDF and LDA compression techniques, decreasing approximately with the square root of the average document size $|d|$ and remaining within 20% of the theoretical upper boundary (45).

5 Discussion

This paper introduces a novel approach to address dimensionality concerns in Hyperdimensional Computing (HDC) by adding compression. We propose a model that combines TF-IDF or LDA-based compression with binary HDC to mitigate the curse of dimensionality. Section 3 presents the core concepts, and Section 4 provides experimental results validating our approach. Our method demonstrates that significantly reducing the encoding space of the initial dictionary only slightly compromises class distinguishability in classification tasks. Specifically, reducing the dictionary by 10 times increases the distinguishability parameter by 10%, and reducing it by 100 times increases the parameter by 40%, while still maintaining a low value (far from 1).

Theorems 1 and 3 provide estimators that accurately estimate the necessary parameters, with TF-IDF compression showing particularly low error and LDA offering slightly better explainability in encoding analysis (see Table 2).

Despite our numerical results aligning with theory, we identify two drawbacks that warrant further research and development in this field:

1. We observe that the bounds provided in Theorems 2–4 are not sufficiently tight. Because these bounds are estimated using the distribution properties of the datasets, it is difficult to obtain tighter bounds for the given metrics.

2. The encoding effect diminishes with increasing vector space size D. This effect, explained by Theorem 3.2, is due to the upper boundary function F becoming concentrated in a narrow region near the line $\varepsilon + 1 = 2\gamma$ as D increases, which reduces the confidence intervals of the estimates $\widehat{\varepsilon}$, without lowering the estimates themselves.

Our results provide several insights into the application of TF-IDF- and LDA-based compression techniques and demonstrate the potential of Compression HDC for broader practical application to empirical problems, where noise significantly hinders data compression and classification.

References

Akiko Aizawa. 2003. An information-theoretic perspective of tf—idf measures. *Inf. Process. Manage.*, 39(1):45–65.

Barry C. Arnold and Richard A. Groeneveld. 1979. Bounds on expectations of linear systematic statistics based on dependent samples. *Annals of Statistics*, 7:220–223.

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

Herman Chernoff. 1952. A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations. *The Annals of Mathematical Statistics*, 23(4):493 – 507.

Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. 2019. On the use of arxiv as a dataset.

P Kanerva. 1988. *Sparse Distributed Memory*. MIT Press, Cambridge, MA.

Pentti Kanerva. 2009. Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive Computation*, 1(2):139–159.

Denis Kleyko, Dmitri Rachkovskij, Evgeny Osipov, and Abbas Rahimi. 2023. A survey on hyperdimensional computing aka vector symbolic architectures, part ii: Applications, cognitive models, and challenges. *ACM Comput. Surv.*, 55(9).

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, pages 74–81. Association for Computational Linguistics.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Anton Mitrokhin, Peter Sutor, Cornelia Fermüller, and Yiannis Aloimonos. 2019. Learning sensorimotor control with neuromorphic sensors: Toward hyperdimensional active perception. *Science Robotics*, 4:eaaw6736.

Peer Neubert and Stefan Schubert. 2021. Hyperdimensional computing as a framework for systematic aggregation of image descriptors. pages 16933–16942.

Saeid Pourmand, Wyatt D. Whiting, Alireza Aghasi, and Nicholas F. Marshall. 2024. Laplace-hdc: Understanding the geometry of binary hyperdimensional computing. *ArXiv*, abs/2404.10759.

Abbas Rahimi, Pentti Kanerva, and Jan M. Rabaey. 2019. Efficient biosignal processing using hyperdimensional computing: A case study for emg-based hand gesture recognition. *IEEE Transactions on Biomedical Engineering*, 66(11):3192–3203.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Anthony Thomas, Sanjoy Dasgupta, and Tajana Rosing. 2021. A theoretical perspective on hyperdimensional computing. *Journal of Artificial Intelligence Research*, 72:215–249.

Tao Yu, Yichi Zhang, Zhiru Zhang, and Christopher De Sa. 2024. Understanding hyperdimensional computing for parallel single-pass learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Guo Yunhui, Mohsen Imani, Jaeyoung Kang, Sahand Salamat, Justin Morris, Baris Aksanli, Yeseong Kim, and Tajana Rosing. 2021. Hyperrec: Efficient recommender systems with hyperdimensional computing. pages 384–389.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NIPS*.

A Appendix / supplemental material

A.1 Compression analysis

In the given section, we provide the theoretical justification of the analysis provided in the paper before. The first part of the upcoming appendix correspond to the TF-IDF and LDA theories.

A.1.1 TF-IDF part

Lemma 1. From Theorem 3 in (Lin, 1991) we have:

$$\text{JSD}(p||q) \leq \frac{1}{2}V(p,q)$$

Using definition of JSD and Pinsker inequality:

$$\text{JSD}(p||q) = \frac{1}{2}\left[D_{KL}(p||m) + D_{KL}(q||m)\right] \geq$$

$$\frac{1}{4}\left[V(p,m) + V(q,m)\right]$$

Now, we are ready to move to the proofs of the theorems.

Theorem 1-Theorem 2. 1. Follow the definition of CR, we have:

$$\text{CR} = \frac{\frac{1}{N}\sum_j |d'|_j}{\frac{1}{N}\sum_j |d|_j} \simeq \frac{\mathbb{E}|d'|}{\mathbb{E}|d|}$$

Based on the model in Assumption 1 we have:

$$\mathbb{E}|d| = \sum_{w\in\mathcal{W}} \frac{\lambda_w^2}{1+\lambda_w} \tag{47}$$

Notice that after the compression procedure, we leave only the words from $\widehat{\mathcal{W}}_p$; hence, given the (47), we have:

$$\text{CR} \simeq \frac{\sum_{w\in\mathcal{W}_p} g(\lambda_w)}{\sum_{w\in\mathcal{W}} g(\lambda_w)},$$

where $g(x) = \dfrac{x^2}{1+x}$. We obtain the result of the Theorem 1.1 by using the consistent estimator (20) for λ_w and using Slutsky's theorem.

Also, we easily obtain the bounds for Theorem 2.1 for $\widehat{\text{CR}}$:

$$\left(pM\frac{\min_{w\in\mathcal{W}_p} g(\widehat{\lambda}_w)}{\sum_w g(\widehat{\lambda}_w)}, pM\frac{\max_{w\in\mathcal{W}_p} g(\widehat{\lambda}_w)}{\sum_w g(\widehat{\lambda}_w)}\right) \tag{48}$$

2. Using the Jensen-Shannon divergence definition and Lemma 3 we have:

$$\text{JSD}(p||q) = \frac{1}{2}\left[D_{KL}(p||m) + D_{KL}(q||m)\right],$$

where $p = \{p_w\}$ and $q = \{q_w\}$, defined in (5). Notice that based on Assumption 1we have the following form for p_w and q_w:

$$p_w = \frac{n_w}{|d|}, \quad q_w = \frac{n_w}{|d'|} \tag{49}$$

Hence we have CR$p_w = q_w$. Next, we can easily find the consistent estimator for p_w:

$$\widehat{p}_w = \frac{\widehat{\lambda}_w}{\sum_k g(\widehat{\lambda_k})}, \tag{50}$$

because of Slutsky's theorem and consistent estimator for λ_w. Now, using the definition of D_{KL}:

$$D_{KL}(p||q) = \sum_w p_w \log\left(\frac{p_w}{q_w}\right), \tag{51}$$

and previous properties: CR$\times p_w = q_w$ and $q_w = 0$ for $w \in \mathcal{W} \setminus \widehat{\mathcal{W}}_p$ we obtain the results.

For the bounds in Theorem 2 we use Lemma 3.

3. ROUGE-L score. Here, we focus on the classical text compression score. ROUGE-L has three components to analyze:

1. Precision: $P = \dfrac{\mathbb{E}|LCS|}{\mathbb{E}_q|d|}$
2. Recall: $R = \dfrac{\mathbb{E}|LCS|}{\mathbb{E}_p|d|}$
3. F-score: $F1 = 2\dfrac{R\cdot P}{R+P}$

Notice that our procedure preserves the order, hence $\mathbb{E}|LCS| = \mathbb{E}_q|d|$. Hence, we have the following:

1. Precision: $P \equiv 1$
2. Recall: $R = \text{CR}$
3. F-score: $F1 = 2\dfrac{\text{CR}}{\text{CR}+1}$

Now, since $f(x) = \dfrac{x}{x+1}$ is increasing for $x \geq 0$, we proved our bounds.

A.1.2 LDA part

Theorem 3-4.

1. Notice that CR $\simeq \dfrac{\mathbb{E}|d'|}{\mathbb{E}|d|}$, hence using Assumption 6

$$\mathbb{E}|d| = \sum_{i=1}^{M}\mathbb{E}v_i = \sum_{i=1}^{M}\sum_{t=1}^{T}\pi_t C\Phi_{t,i},$$

where π_t - probability of document's topic is t. Hence using the

$$\widehat{\text{CR}} = \frac{\sum_{i=1}^{M}\sum_{w\in\mathcal{W}_{t,tp}}\widehat{\pi}_t\Phi_{t,w}}{\sum_{i=1}^{M}\sum_{w}\widehat{\pi}_t\Phi_{t,w}},$$

where $\widehat{\pi}_t = \dfrac{1}{N}\sum_{j=1}^{N} z_{t,d_j}$ we obtain the consistent estimator of the CR.

The upper bound can be obtained as follows:

$$\widehat{\text{CR}} = \sum_t \pi_t \sum_{w\in\mathcal{W}_{p,t}} \Phi_{w,t},$$

where $\Phi_{w,t} \approx \mathbb{E}X_{(j)}$, j corresponding number of order statistics and $X = \{X_1, \dots, X_M\}$ sequence of Beta distributed r.v. as in 2. Hence using the $\sum_t \pi_t = 1$, we can proceed with the Lemma 3 to obtain:

$$\widehat{\text{CR}} \geq p - \frac{\sqrt{2}}{M}\sum_{i=\lceil(1-p)M\rceil}^{M}\sqrt{\frac{M-i}{i}} \tag{52}$$

$$\widehat{\text{CR}} \leq \left(p + \frac{\sqrt{2}}{M}\sum_{i=\lceil(1-p)M\rceil}^{M}\sqrt{\frac{i-1}{M-i+1}}\right) \tag{53}$$

This leads us to the following:

$$p - \sqrt{\frac{2p}{1-p}} \leq \widehat{\text{CR}} \leq p + p\sqrt{2(M-1)}$$

2. We want to examine the value of the:

$$\overline{\text{JSD}(p||q)} = \sum_{t=1}^{T}\pi_t \text{JSD}(p_t, q_t),$$

where $p_{t,i} = \dfrac{1}{N_t}\sum_{j=1}^{N} f_{i,j}z_{t,d_j}$ and $q_{t,i} = \dfrac{1}{N_t}\sum_{j=1}^{N} f'_{i,j}z_{t,d_j}$. Under assumption Assumption 6, we have:

$$p_{t,i}/q_{t,i} = f_i/f'_i = 1/\text{CR}$$

Therefore, we have: $\text{CR}\times p_{t,i} = q_{t,i}$. Also, we have:

$$\widehat{p}_{t,i} = \frac{C\times\Phi_{t,i}}{\sum_k C\times\Phi_{t,k}} = \Phi_{t,i} \xrightarrow{\mathbb{P}} p_{t,i},$$

hence using Slutsky's theorem and consistent estimators for π_t and $p_{t,i}$, $q_{t,i}$ we have the consistent estimator.

Bounds for JSD are obtained as in the proof of Theorem 2, using the definition (6)

3. The same idea as in the proof of the Theorem 2 works here.

A.2 Encoding analysis

In the given section, we provide the theoretical justification of the encoding analysis, provided in the paper.

Lemma 4. Let's consider $u_1, \dots, u_D$ independent uniform distributions on $[0, 1]$. Denote $Y_i = \mathbf{1}(u_i \leq p)$, then $\{Y_i\}_i$ are independent. Here we assume $\mathbf{1}(...) \in \{\pm 1\}$, to satisfy the Bernoulli-type distribution of X_i.

Notice that $\mathbb{P}(X_i = 1) = \mathbb{P}(u_i \leq q_i)$, where $q_i = \mathbb{P}(X_i = 1|X_1, \dots, X_{i-1})$ and thence:

$$X_i \leq Y_i \Rightarrow \mathbb{P}(\sum_{i=1}^{D} X_i \geq \varepsilon D) \leq \mathbb{P}\left(\sum_{i=1}^{D} Y_i \geq \varepsilon D\right)$$

Theorem 5.

Probability estimation part.

In the given appendix, we justify the ideas provided in the **encoding** part in the theory section. Notice that we aimed to consider the given probability:

$$\mathbb{P}(\langle\phi(d_1), \phi(d_2)\rangle \geq \varepsilon D) = \mathbb{P}\left(\sum_{i=1}^{D} X_i \geq \varepsilon D\right) = \bigstar$$

Using the Lemma 5, we can obtain:

$$\bigstar \leq \inf_{t>0}\left[e^{-\varepsilon D t}\left(\mathbb{E}e^{tX}\right)^D\right],$$

where X is a Bernoulli random variable with parameter γ and values in $\{\pm 1\}$. Hence, we have:

$$\bigstar \leq \inf_{t>0}\left[e^{-\varepsilon D t}\left(\gamma e^t + (1-\gamma)e^{-t}\right)^D\right] = \inf_{t>0} L(t)$$

To find the minimum of the $L(t)$, we need to derive the first-order condition:

$$\frac{\mathrm{d}}{\mathrm{d}t}L(t) = 0$$

This is equivalent to:

$$\underbrace{(\gamma(e^{2t}-1)+1)^{D-1}}_{>0,\text{ since }\gamma<1} \times \left((\gamma-1)(\varepsilon D + D) - \gamma e^{2t}(\varepsilon D - D)\right) = 0$$

$$(1-\gamma)D(\varepsilon+1) = \gamma D(1-\varepsilon)e^{2t} \Rightarrow$$

$$t_{\min} = \frac{1}{2}\ln\left[\frac{1-\gamma}{\gamma}\frac{1+\varepsilon}{1-\varepsilon}\right] = \frac{1}{2}\ln\underbrace{C(\varepsilon)C(\gamma)}_{C(\varepsilon,\gamma)}$$

After rearranging, we have:

$$\exp\Big[-D\Big(\varepsilon\ln\sqrt{C(\varepsilon,\gamma)}$$

$$-\ln\Big(p\sqrt{C(\varepsilon,\gamma)}+\tfrac{1-\gamma}{\sqrt{C(\varepsilon,\gamma)}}\Big)\Big)\Big] =$$

$$\exp\left[-D\ln\left(\frac{C(\varepsilon,\gamma)^{(\varepsilon+1)/2}}{1-\gamma+\gamma C(\varepsilon,\gamma)}\right)\right] =$$

$$\exp\left[-D\ln\underbrace{\left(\frac{1}{2}\left[\frac{1-\gamma}{\gamma}\frac{1+\varepsilon}{1-\varepsilon}\right]^{(\varepsilon-1)/2}\frac{1+\varepsilon}{\gamma}\right)}_{\star\star}\right]$$

Hence, this probability decreases with increasing D or by managing the expression in scopes. Simple algebra shows that for the same level of D and ε, we can increase the expression $\star\star$ by increasing the γ value after the critical point $\gamma_\varepsilon = \dfrac{1+\varepsilon}{2}$.

Compression connection part.

Next, we aim to connect the encoding analysis with the compression part. We provide the following explanation. Consider the following relationship:

$$\gamma = \mathbb{P}(\phi_{1,i}\phi_{2,i} = 1) = \tilde{\gamma}^2 + (1-\tilde{\gamma})^2$$

where ϕ_i is the i-th position of the vector-encoding of randomly generated document d.

Notice that:

$$\tilde{\gamma} = \mathbb{P}\left(\operatorname{sign}\left[\sum_{j=1}^{|d|}\phi_{i,j}\right] = 1\right) =$$

$$\mathbb{P}\left(\operatorname{sign}\left[\underbrace{\sum_{k=1}^{M}\#\{w_k\}\phi_{i,w_k}}_{\nu_i}\right] = 1\right),$$

where the support of the ν_i is determined by the all possible sums of $\sum_{k=1}^{M}\pm\#\{w_k\}$. The behavior of this sum is quite unpredictable, but we can say that the given distribution is symmetrical. To estimate $\mathbb{P}(\operatorname{sign}\nu_i = 1)$ we will consider the probability of $\eta = \mathbb{P}(\nu_i = 0)$. Hence (by symmetry), we have:

$$\tilde{\gamma} = \frac{1}{2} + \frac{\eta}{2},$$

i.e., we cut half of the probability from the left tail of the distribution and add it to the right one. We propose the following estimation of the η:

$$\eta \le \binom{|d|}{\lceil |d|/2\rceil}\frac{1}{2^{|d|}}$$

This bound is easy to obtain assuming $\nu_i \approx \sum_{i=1}^{|d|}\upsilon_i$, where υ_i is independent Bernoulli r.v. with values ± 1 and equal proabilities.

Based on the CR definition, $\mathrm{CR}\times|d| = |d'|$, hence for compressed object the value of η will be bounded by:

$$\eta \le \binom{\mathrm{CR}\,|d|}{\lceil \mathrm{CR}\,|d|/2\rceil}\frac{1}{2^{\mathrm{CR}\,|d|}}$$

The RHS is increasing with the decreasing of the CR. As a result, we have:

$$\gamma = \tilde{\gamma}^2 + (1-\tilde{\gamma})^2 = \left(\frac{1}{2}+\frac{\eta}{2}\right)^2 + \left(\frac{1}{2}-\frac{\eta}{2}\right)^2 \le$$

$$\frac{1}{2} + \binom{|d|}{\lceil |d|/2\rceil}\frac{1}{2^{|d|}}$$

A.3 Additional results

In the given section we provide the figures, providing a comprehensive compression analysis comparing TF-IDF and LDA techniques across three distinct datasets (IMDB, AG News, and arXiv). The analysis evaluates three key metrics - Compression Ratio (CR), Jensen-Shannon Divergence (JSD), and ROUGE-F1 scores - as functions of dictionary compression quantile p, with results plotted against their theoretical estimators. The green shaded regions represent confidence intervals around the estimated values, while the black dots indicate the true theoretical values for comparison. Both TF-IDF (top three rows) and LDA (bottom three rows) methods show varying performance patterns across the different datasets, with the estimation curves generally tracking well with their corresponding theoretical benchmarks.

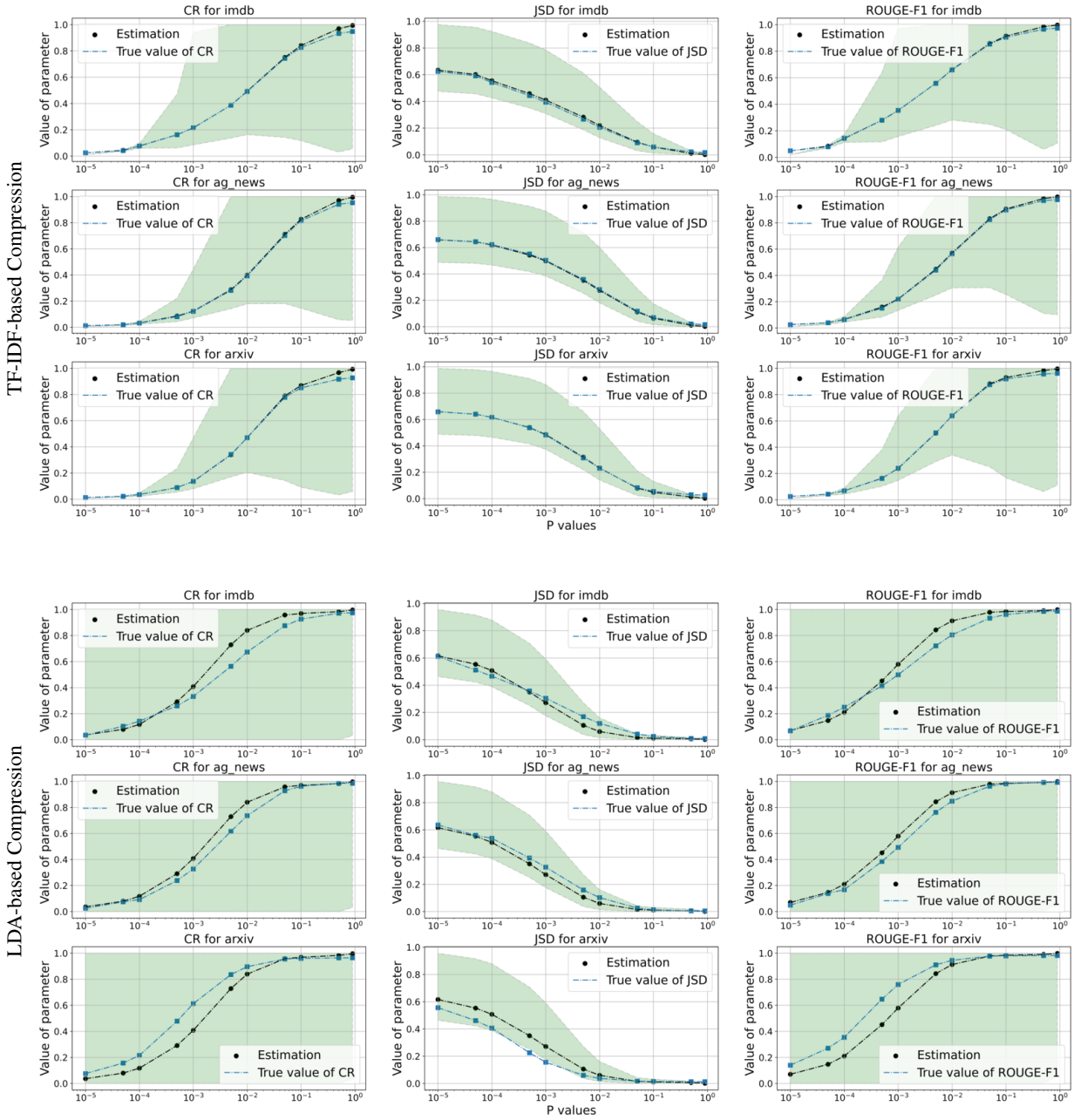

Figure 3: Compression analysis for TF-IDF (top three rows) and LDA (bottom three rows) techniques. The results compare the compression ratio CR, Jensen-Shannon divergence JSD, and ROUGE-F1 scores, as functions of the dictionary compression quantile p, with their theoretical estimators across the IMDB, AG News, and arXiv datasets .

A.4 Experiment algorithms

Here, we describe the algorithms referenced in the main text and used throughout the experimental section. For both of the central components of the paper – the analysis of compression-based representations and the evaluation of statistical bounds – we provide clear pseudo-code that can be directly translated into practical implementations. The goal of presenting the algorithms in the appendix is to give the reader a transparent view of how the theoretical quantities are computed in practice, bridging the gap between abstract definitions and experimental procedures. Each algorithm is written in a way that emphasizes the logical flow of operations, starting from the input dataset, applying compression or transformation, and proceeding to the estimation of key quantities such as divergences, bounds, and error measures. By doing so, we aim to highlight that the computational steps are straightforward and reproducible, and that they can be adapted to other datasets or models with minimal modification.

Algorithm 1 Clusterization statistics collection

Input: Dataset X, compression model $f_{comp} \in \{\text{tf-idf}, \text{LDA}\}$, p_{values} list of possible compression parameters.
Return: D_p dictionary of statistics.
$D_p \leftarrow \{\}$
for p in p_{values} **do**
  $X_c \leftarrow f_{comp}(X, p)$
  $\widehat{Y}_p \leftarrow Stats(X_c, p)$ {Calculate statistics based on Theorems 1 – 4 with X_c}
  $Y_p \leftarrow TrueValues(X_c, p)$ {Calculate true values based on definitions in Section 3.1.}
  $D_p[p] \leftarrow (\widehat{Y}_p, Y_p)$ {Save the bounds and estimators for the given value of p}
end for

Algorithm 2 Encoding statistics collection

Input: Dataset X, dimension size D, $epochs$ number of epochs of Monte Carlp, compression model $f_{comp} \in \{\text{tf-idf}, \text{LDA}\}$, p_{values} list of possible compression parameters.
Return: E the list of encoding statistics
$E \leftarrow []$
for i in $[1, \ldots, epochs]$ **do**
  $\Phi(\mathcal{A}) \leftarrow U(\{\pm 1\}^{|\mathcal{A} \times d|})$ {Generate random vectors}
  $\widehat{\varepsilon_p} \leftarrow \{p : 0\}$ {Dict for interesting values of p}
  for j in $[1, \ldots, epochs]$ **do**
    for p in p_{values} **do**
      $d'_1, d'_2 \leftarrow f_{comp}(d_1, p), f_{comp}(d_2, p)$ {Compress the documents}
      $\phi'_1, \phi'_2 \leftarrow \phi(d'_1), \phi(d'_2)$ {Encode the documents}
      $\widehat{\varepsilon}_p[p] = \widehat{\varepsilon}_p[p] + \dfrac{|\langle \phi'_1, \phi'_2 \rangle|}{D}$
    end for
  end for
  $\widehat{\varepsilon}_p[p] = \widehat{\varepsilon}_p[p]/epochs$ {Average the value of dot-product}
  $E = E \cup \widehat{\varepsilon}_p$
end for
$E = (\text{mean}(E), \text{std}(E))$ {Average and get std of all estimators}