# ONE SWALLOW DOES NOT MAKE A SUMMER: UNDERSTANDING SEMANTIC STRUCTURES IN EMBEDDING SPACES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The geometry of local embedding neighborhoods provides context for interpreting their members. We investigate how changes in this geometry as neighborhoods expand reveal semantic structure encoded in the embedding space. We introduce the **Semantic Field Subspace (SFS)** to summarize the geometric structure of a local neighborhood while preserving the original embedding geometry. We then introduce **Semantic Shift**, which characterizes neighborhood evolution through changes in subspace directions and their relative importance. **SAFARI** (**S**em**A**ntic **F**ield subsp**A**ce dete**R**m**I**nation) is an unsupervised algorithm that tracks Semantic Shift during hierarchical clustering and selects SFSes at significant transitions, revealing hierarchical relationships among semantic fields. An efficient approximation of Semantic Shift achieves a 15–30$\times$ speedup over full SVD, with average errors below 0.01. Experiments on five text datasets show that **SFSes** outperform standard classifiers in text classification and political bias detection, while **SAFARI** reveals interpretable semantic hierarchies. Additional experiments on image embeddings demonstrate applicability across modalities. This perspective links local semantic context to hierarchical organization through neighborhood evolution.

## 1 INTRODUCTION

Embedding spaces are foundational to modern natural language processing, converting text into dense vectors that capture semantic properties in a tractable format. By translating semantic similarity into geometric proximity, these spaces enable efficient comparison, retrieval, and manipulation. This property supports a wide range of applications, including knowledge retrieval (Lewis et al., 2020; Guu et al., 2020; Ram et al., 2023; Asai et al., 2024), personalized and diverse recommendations (Gan et al., 2020; Hirata et al., 2022; Huang et al., 2024; Sun et al., 2024), and multimodal understanding (Yu et al., 2019; Luo et al., 2023; Zhang et al., 2024; Yu et al., 2023). Yet, despite their centrality, embedding spaces are often treated as black boxes, limiting interpretability and constraining targeted adaptation for downstream tasks.

Research on understanding embedding spaces generally falls into two directions. The first analyzes geometric properties to enhance both representational quality and interpretability. In Natural Language Processing (NLP), post-processing methods have tackled issues such as anisotropy and instability (Mu & Viswanath, 2018; Liu et al., 2019), while studies on contextual embeddings reveal expressivity limits (Ethayarajh, 2019). Techniques like rotation-based alignment and probing further link embedding dimensions to interpretable concepts (Park et al., 2017; Dufter & Schütze, 2019; Dalvi et al., 2019; Clark et al., 2019). Similar geometric challenges, e.g., feature collapse and variance concentration, are found in visual (Chen et al., 2020; He et al., 2020; Grill et al., 2020) and multimodal embeddings (Radford et al., 2021; Jia et al., 2021).

The second direction explores latent semantic and hierarchical structures within embedding spaces. In NLP, embeddings have been mapped to external conceptual systems for flexible semantic interpretations (Simhi & Markovitch, 2023), while in vision, researchers have identified neurons encoding abstract, multimodal concepts (Goh et al., 2021). Unsupervised clustering (Van Gansbeke et al., 2020; Caron et al., 2020) uncover semantic groupings, and hierarchical classification approaches (Deng et al., 2012; Dhall et al., 2020) organize semantics into multi-level taxonomies.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Connecting geometric analysis with semantic structure discovery calls for understanding how local embedding neighborhoods evolve. This perspective raises three challenges.

- **Representing semantic context.** An embedding's meaning depends on the related embeddings that supply its context. Capturing this context requires a representation of local neighborhood geometry that preserves the structure of the original embedding space.
- **Identifying semantic transitions.** As a neighborhood expands, added embeddings can reinforce an existing concept or introduce broader semantic context. Distinguishing these changes requires a geometric criterion for detecting meaningful transitions.
- **Connecting semantic scales.** Local neighborhoods capture semantic relationships at different levels of granularity. Organizing these relationships into a hierarchy requires tracing how semantic fields relate as their neighborhoods expand.

We investigate how the expansion of local embedding neighborhoods reveals semantic structure encoded in their native geometry. A neighborhood provides context for interpreting its members, and changes in its geometric representation offer signals of semantic transitions as that context broadens. This perspective connects local semantic representation, semantic transition, and hierarchical structure discovery. Our contributions are as follows.

- **Semantic representation.** We introduce Semantic Field Subspaces (SFSes) to represent the semantic context supplied by local neighborhoods. SFSes summarize the geometric structure of related embeddings, providing a context-aware representation that preserves the original embedding geometry.
- **Semantic transition.** We introduce Semantic Shift to characterize how neighborhood representations change during expansion. By combining changes in subspace directions and their relative importance, Semantic Shift provides a geometric criterion for identifying boundaries between semantic fields.
- **Structure discovery.** We propose SAFARI, an unsupervised algorithm that tracks Semantic Shift during hierarchical clustering and selects SFSes at significant transitions. The resulting relationships organize semantic fields across scales within the original embedding space. An efficient approximation of Semantic Shift supports scalable discovery.

Experiments on five text datasets, complemented by image embedding experiments, demonstrate the framework's applicability across modalities. SFSes outperform standard classifiers on text classification and capture fine-grained distinctions in political bias, while SAFARI reveals hierarchical relationships among local semantic fields. Our Semantic Shift approximation achieves a 15–30$\times$ speedup over full SVD, with average errors below 0.01.

## 2 Related Work

Research on understanding embedding spaces generally follows two directions: (1) structural analysis of geometric properties and (2) discovery of latent semantic hierarchies.

**Structural Analysis of Embedding Spaces.** Early work focuses on geometric properties of embeddings. For example, Mu & Viswanath (2018) mitigated anisotropy by removing dominant principal components, while Liu et al. (2019) stabilized embedding distributions by suppressing high-variance dimensions. Ethayarajh (2019) found that contextualized embeddings often cluster in narrow cones, limiting expressiveness. To improve interpretability, rotation-based alignment (Park et al., 2017; Dufter & Schütze, 2019) and probing techniques (Dalvi et al., 2019; Clark et al., 2019) linked embedding dimensions to human-understandable concepts.

Similar geometric issues, such as feature collapse and variance concentration, also arise in visual representations from self-supervised models like SimCLR (Chen et al., 2020), MoCo (He et al., 2020), and BYOL (Grill et al., 2020). Wang & Isola (2020) analyzed such effects in contrastive learning, while feature visualization methods (Olah et al., 2017; Zhou et al., 2016) offered neuron-level insights. Recent advances in multimodal embeddings, e.g., CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and DeCLIP (Li et al., 2022), explore joint semantic alignment across modalities. Unlike these approaches, which often modify embedding spaces, **SAFARI** preserves native geometry while enhancing interpretability.

**Semantic and Hierarchical Structure Discovery.** A complementary line of research seeks to uncover semantic groupings and hierarchies within embedding spaces. In NLP, Simhi & Markovitch (2023) maps embeddings into conceptual spaces grounded in knowledge bases, enabling flexible interpretations. In vision, Goh et al. (2021) found that certain CLIP neurons respond to abstract concepts shared across modalities. Unsupervised methods (Van Gansbeke et al., 2020; Caron et al., 2020) discover coherent visual clusters but lack hierarchical organization. Conversely, hierarchical classification methods (Deng et al., 2012; Dhall et al., 2020) build multi-level structures using taxonomies like WordNet, yet depend on external supervision and predefined label trees. **SAFARI** bridges these gaps with a unified, unsupervised framework that identifies semantic hierarchies directly from embedding spaces–without structural transformation or supervision.

## 3 Problem Formulation

**Vector Space Foundation.** We model embedding spaces as a vector space $\mathbb{R}^d$. For ease of explaining the core concepts in this paper, we use natural language terms as illustrative examples, though all concepts introduced generalize beyond text. Let $h : \mathcal{T} \to \mathcal{E}$ be a deep model that maps real-world terms $\mathcal{T}$ to embedding vectors $\mathcal{E}$. A central assumption is that **geometric distances reflect semantic similarities**, a principle validated in various tasks (Karpukhin et al., 2020; Lewis et al., 2020; Ram et al., 2023). In this work, we adopt cosine distance as a proxy for semantic dissimilarity.

**Definition 3.1** (Semantic Distance)**:** *The semantic distance $d_{sem}(\cdot, \cdot)$ between two embedding vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ is defined as $d_{sem}(\boldsymbol{u}, \boldsymbol{v}) = 1 - \frac{\langle \boldsymbol{u}, \boldsymbol{v} \rangle}{\|\boldsymbol{u}\| \|\boldsymbol{v}\|}$.*

**Challenge of Context-Dependent Meaning.** Despite precise embeddings, **semantic meaning depends on context**. Linguistic theories such as semantic field theory and componential analysis (Ullmann, 1957; Nida, 2015) argue that meaning arises only through contextual associations.

**Proposition 3.1** (Context-Dependent Meaning)**:** *An embedding vector cannot be semantically interpreted in isolation.*

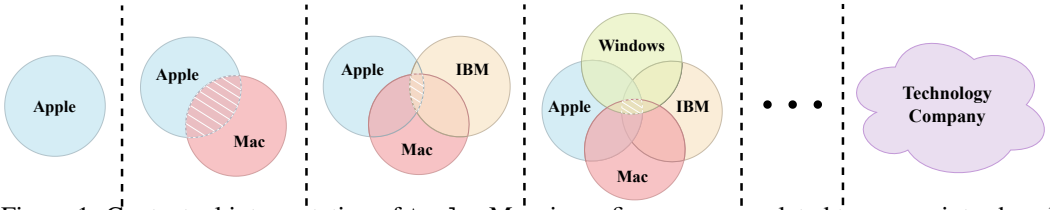



Figure 1: Contextual interpretation of `Apple`: Meaning refines as more related terms are introduced.

**Example 3.1:** The term `Apple` is semantically ambiguous and relies on context for disambiguation. As shown in Figure 1, it refers to a tech company when grouped with `Mac`, `IBM`, and `Windows`, `Apple`, but to a fruit with `Apple Tree`, `Juice`, and `Banana`. More contextual cues yield more precise interpretations, highlighting the context-dependent nature of semantics in embedding spaces. △

**Semantic Fields in Embedding Spaces.** To model context-dependent semantics, we define **Semantic Fields** as sets of neighboring vectors that contextualize a target embedding vector. For instance, as shown in Example 3.1, the meaning of `Apple` becomes clearer when surrounded by neighbors like `Mac`, `IBM`, and `Windows`, forming a Semantic Field of `Apple`. This concept aligns with foundational embedding models such as Word2Vec (Mikolov et al., 2013) and BERT (Devlin et al., 2019), where semantics arise from context. To formalize this, we distinguish general embedding vectors ($\boldsymbol{v}$) from those representing real-world terms ($\boldsymbol{v}_t$), where the latter excludes purely mathematical constructs (e.g., zero vectors).

**Definition 3.2** (Semantic Field)**:** *A set $\mathcal{F}$ of embedding vectors forms a Semantic Field of radius $\epsilon > 0$ if there exists a central vector $\boldsymbol{v}_t \in \mathcal{F}$ such that for all $\boldsymbol{u}_t \in \mathcal{F}$, we have $d_{sem}(\boldsymbol{u}_t, \boldsymbol{v}_t) < \epsilon$.*

While Definition 3.2 allows us to examine local structures in an embedding space, it is insufficient for understanding the global organization of semantics.

**Research Objective: From Local to Global Semantics.** Our goal is to investigate how Semantic Fields collectively shape the global semantic structure of embedding spaces, uncovering how these

local Semantic Fields relate, interact, and form coherent semantic hierarchies. To this end, we propose **SAFARI**, a principled framework that detects and analyzes hierarchical semantic structures by identifying boundaries between Semantic Fields. This method offers an interpretable lens for understanding how semantics are organized in high-dimensional embedding spaces.

# 4 METHODOLOGY

## 4.1 SEMANTIC FIELD REPRESENTATION

By Definition 3.2, a Semantic Field is the neighborhood of a target embedding vector. **SAFARI** identifies the structure of these neighborhoods. However, representing such structures is non-trivial due to the following two key challenges:

- **Handling Expression Variants:** Closest neighbors are often populated by variants that are overly similar to the original, offering limited values into its interpretation (Mimno & Thompson, 2017; Ethayarajh, 2019). For example, word synonyms or images of the same object under different lighting conditions contribute little to understanding the underlying concept. A robust representation should exhibit invariance to such variations.
- **Delineating Semantic Field Boundaries:** As a Semantic Field expands, it includes more embedding vectors in larger neighborhoods and thus provides richer context for more refined interpretations of the target embedding. However, since semantics naturally form hierarchies, there exists a boundary beyond which the Semantic Field contains embeddings diverse enough for it to transcend from a concrete concept to a more abstract one. It is necessary to determine such boundaries to tell when the Semantic Field starts to represent a new concept.

As illustrated in Figure 2, starting from `Coca-Cola`, the nearest neighbor is `Coke` (a lexical variant), while broader semantic context only occurs with terms like `Sprite` and `Pepsi`, underscoring the difficulty of managing expression variants and identifying natural semantic boundaries.

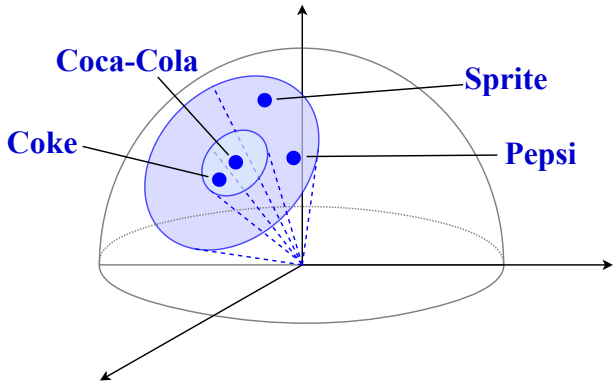



Figure 2: Illustration of Semantic Field exploration.

**Geometric Representation: Semantic Field Subspace (SFS).** To address these challenges, we introduce the Semantic Field Subspaces (SFS), a low-dimensional subspace spanned by semantically related vectors. This geometric representation naturally absorbs expression variants via linear dependence and provides a compact, geometry-preserving abstraction of semantic content.

**Definition 4.1** (Semantic Field Subspace (SFS))**:** *Let $\mathcal{F} = \{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n\}$ be a Semantic Field. Its SFS is defined as:* $\mathbb{S}_{\mathcal{F}} = span(\mathcal{F}) = \{\sum_{i=1}^{n} \alpha_i \boldsymbol{v}_i \mid \alpha_i \in \mathbb{R}\}$.

We compute the basis of $\mathbb{S}_{\mathcal{F}}$ via SVD on matrix $\boldsymbol{M}_{\mathcal{F}} = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n]$, i.e., $\boldsymbol{M}_{\mathcal{F}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$. In practice, since real-world embedding matrices are noisy and numerically full-rank, we do not use the literal span but represent $\mathbb{S}_{\mathcal{F}}$ by its singular directions weighted by their singular values, optionally truncated to the leading ones; this form is used in Semantic Shift (Eq. 1) and all experiments (settings in Appendix A.3). This yields a continuous, low-rank representation that captures semantic structure while remaining invariant to redundancy.

**Algorithm 1: SAFARI**

**Input:** Embedding set $\mathcal{E} \subset \mathbb{R}^d$, window size $w$;
**Output:** Set $\Phi$ of Semantic Field Subspaces (SFSes);

$\Omega \leftarrow$ Initialize each $\boldsymbol{v}_t \in \mathcal{E}$ as its own cluster;
$\mu \leftarrow 0$; $\tau \leftarrow 0$; $\Phi \leftarrow \varnothing$;
**while** $|\Omega| > 1$ **do**
  ▷ `Step 1: Cluster Merging`
  $\{\mathcal{C}_x, \mathcal{C}_y\} \leftarrow \arg\min_{\mathcal{C}_i, \mathcal{C}_j \in \Omega} d_{sem}(\mathcal{C}_i, \mathcal{C}_j)$;
  $\mathcal{C}_{new} \leftarrow \mathcal{C}_x \cup \mathcal{C}_y$;
  $\Omega \leftarrow \Omega \cup \{\mathcal{C}_{new}\} \setminus \{\mathcal{C}_x, \mathcal{C}_y\}$;
  ▷ `Step 2: SFS Delineation`
  $\mathcal{C}_x \leftarrow |\mathcal{C}_x| > |\mathcal{C}_y| \; ? \; \mathcal{C}_x \; : \; \mathcal{C}_y$;
  $\mathbb{S}_x, \mathbb{S}_{new} \leftarrow \text{span}(\mathcal{C}_x), \text{span}(\mathcal{C}_{new})$;
  Compute $\Delta F_{sem}(\mathbb{S}_x, \mathbb{S}_{new})$ using Eq. 1;
  **if** $\Delta F_{sem}(\mathbb{S}_x, \mathbb{S}_{new}) > \mu + 3\tau$ **then**
    $\Phi \leftarrow \Phi \cup \{\mathbb{S}_{new}\}$;
  Recompute the sliding-window mean $\mu$ and standard deviation $\tau$ from the last $w$ Semantic Shift values, updating the dynamic threshold for the next iteration;
**return** $\Phi$;

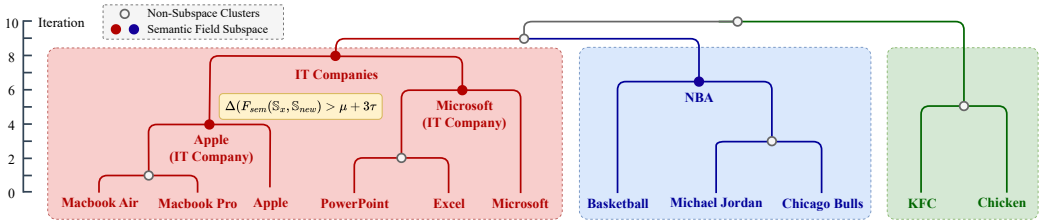



Figure 3: Toy example illustrating **SAFARI**'s hierarchical clustering process.

**SFS Boundary Delineation via Semantic Shift.** To delineate boundaries between Semantic Fields, we leverage the hierarchical nature of language semantics, where expanding fields lead to broader, more abstract concepts.

**Proposition 4.1** (Hierarchical Semantic Structure)**:** *Semantic hierarchies in natural language are reflected in the geometric structure of embedding spaces.*

Importantly, the hierarchies discovered within embedding spaces are **native** to those spaces, shaped by the model architecture, training objective, and data distribution. They should not be expected to mirror human-designed ontologies; rather, they reveal how the model organizes semantics, which may diverge from human intuition (see Appendix E for concrete examples).

As a Semantic Field expands, a significant shift in meaning often indicates an evolution to a new field. We quantify this evolution via the **Semantic Shift** between two subspaces $\mathbb{S}_{\mathcal{F}_x}$ and $\mathbb{S}_{\mathcal{F}_{new}}$:

**Definition 4.2** (Semantic Shift)**:** *The Semantic Shift between $\mathbb{S}_{\mathcal{F}_x}$ and $\mathbb{S}_{\mathcal{F}_{new}}$ is defined as:*

$$\Delta F_{sem}(\mathbb{S}_{\mathcal{F}_x}, \mathbb{S}_{\mathcal{F}_{new}}) = \textstyle\sum_i \Delta\sigma_i \cdot d_{sem}(\boldsymbol{v}_i, \tilde{\boldsymbol{v}}_i^*), \tag{1}$$

*where $\Delta\sigma_i = |\sigma_i - \tilde{\sigma}_i|$ captures the **dimensional importance shift** in singular values $\sigma_i \in \boldsymbol{\Sigma}_x$ and $\tilde{\sigma}_i \in \boldsymbol{\Sigma}_{new}$; $d_{sem}(\boldsymbol{v}_i, \tilde{\boldsymbol{v}}_i^*)$ captures **directional change** between basis vectors $\boldsymbol{v}_i \in \boldsymbol{V}_x^\top$ and their nearest counterparts $\tilde{\boldsymbol{v}}_i^* \in \boldsymbol{V}_{new}^\top$.*

Semantic Shift acts as a boundary criterion: A large shift suggests that the new subspace $\mathbb{S}_{\mathcal{F}_{new}}$ represents a more abstract concept that subsumes $\mathbb{S}_{\mathcal{F}_x}$, whereas small values reflect refinements within the same semantic field.

### 4.2 The SAFARI Algorithm

**Algorithm Overview.** Building on Definitions 4.1 and 4.2, we propose **SAFARI**, an algorithm to uncover SFSes by monitoring Semantic Shifts during iterative clustering. At each step, **SAFARI**

merges the nearest clusters, resulting in a new subspace. The algorithm then evaluates the Semantic Shift between the new subspace and the previous subspace and checks whether such a shift is significant. If so, the new subspace is identified as a new SFS that subsumes the previous subspaces.

**Detailed Procedure.** The pseudo-code is provided in Algorithm 1. **SAFARI** initializes by assigning each vector as a singleton cluster in a set $\Omega$, and maintains a set $\Phi$ to store the discovered SFSes. It proceeds iteratively with the steps below until only one cluster remains ($|\Omega| \leq 1$):

- **Step 1: Cluster Merging.** The two nearest clusters $\mathcal{C}_x$ and $\mathcal{C}_y$ are identified using Semantic Distance $d_{sem}(\mathcal{C}_x, \mathcal{C}_y)$, with centroids representing each cluster. They are then merged into a new cluster $\mathcal{C}_{new}$, after which $\mathcal{C}_x$ and $\mathcal{C}_y$ are removed, and $\mathcal{C}_{new}$ is added to $\Omega$.
- **Step 2: SFS Delineation. SAFARI** constructs the subspaces $\mathbb{S}_{new}$ and $\mathbb{S}_x$ for $\mathcal{C}_{new}$ and the larger cluster $\mathcal{C}_x$, and computes Semantic Shift $\Delta F_{sem}(\mathbb{S}_x, \mathbb{S}_{new})$. A sliding window of size $w$ tracks the recent $w$ values, computing mean $\mu$ and standard deviation $\tau$. If the current $\Delta F_{sem}(\mathbb{S}_x, \mathbb{S}_{new})$ exceeds the dynamic threshold ($\mu + 3\tau$), $\mathbb{S}_{new}$ is added to $\Phi$ as a new SFS.

**Remarks on Design.** The use of a sliding window ($w$) and the dynamic threshold ($\mu + 3\tau$), which is obtained via parameter study, is essential to accurately identify SFSes because we empirically observe that the baseline values of Semantic Shifts grow gradually as the algorithm proceeds. The dynamic threshold allows **SAFARI** to adapt to such gradual change and effectively detect local spikes of real significance. Moreover, **SAFARI** adopts the hierarchical process to reflect the layered nature of semantic relationships: (1) It dynamically determines SFSes based on semantic structures, avoiding relying on pre-defined cluster counts; (2) It reveals natural semantic hierarchies through the resulting dendrogram.

**Example 4.1:** Consider a toy dataset of 11 textual terms. Figure 3 shows how **SAFARI** identifies SFSes. In the first three iterations, semantically close pairs (e.g., `Macbook Air` & `Macbook Pro`, `PowerPoint` & `Excel`, and `Michael Jordan` & `Chicago Bulls`) are merged without forming SFSes. In the 4th iteration, merging `Apple` with the `Macbook` cluster triggers a significant Semantic Shift, forming a new SFS. By the 8th iteration, a hierarchical structure emerges with an `IT Companies` subspace encompassing nested SFSes for `Apple (IT Company)` and `Microsoft (IT Company)`. In the 9th iteration, the dynamic threshold keeps `IT Companies` and `NBA` as distinct SFSes, preserving semantic boundaries. △

### 4.3 Efficient Approximation of Semantic Shift

Each iteration of **SAFARI** requires computing the Semantic Shift via full SVD on matrices of size $n \times d$ ($d \leq n$), incurring a time complexity of $O(nd^2)$ (Halko et al., 2009; Trefethen & Bau, 2022). This becomes a major bottleneck in large-scale applications. To improve scalability, we propose a practical approximation. Let $\boldsymbol{A}_x$ and $\boldsymbol{A}_y$ be the matrices of a large cluster $\mathcal{C}_x$ and a small one $\mathcal{C}_y$. Instead of computing full SVDs, we approximate the Semantic Shift between $\mathbb{S}_x$ and $\mathbb{S}_{new}$ as:

$$\Delta \tilde{F}_{sem}(\mathbb{S}_x, \mathbb{S}_{new}) = \|\boldsymbol{A}_y\|_2 \, \sigma_{max}(\boldsymbol{A}_x), \tag{2}$$

where $\|\boldsymbol{A}_y\|_2$ is the spectral norm of $\boldsymbol{A}_y$, and $\sigma_{max}(\boldsymbol{A}_x)$ is the largest singular value of $\boldsymbol{A}_x$. This yields substantial speedups with negligible loss of accuracy (see Section 5.4).

**Theoretical Justification.** Let $\boldsymbol{A}_{new} = [\boldsymbol{A}_x | \boldsymbol{A}_y]$ be the matrix representing the newly merged cluster $\mathcal{C}_{new}$. We justify the approximation by bounding the dimensional importance shift:

**Theorem 4.1** (Bound on Dimensional Importance Shift)**:** *Given matrices $\boldsymbol{A}_x$ and $\boldsymbol{A}_y$ with the same number of columns and assuming $\boldsymbol{A}_x$ has more rows than $\boldsymbol{A}_y$, the shift in the $i$-th singular value satisfies:*

$$\Delta\sigma_i = |\sigma_i(\boldsymbol{A}_x) - \sigma_i(\boldsymbol{A}_{new})| \leq \|\boldsymbol{A}_y\|_2 .$$

The proof applies Weyl's Theorem (Weyl, 1912) and is provided in Appendix F. The directional change $\sum_i d_{sem}(\boldsymbol{v}_i, \tilde{\boldsymbol{v}}_i^*)$ can further be approximated by $\sigma_{max}(\boldsymbol{A}_x)$ via the condition number of $\boldsymbol{A}_x$ (Meyer, 2000; Belsley et al., 2005); the full derivation is also in Appendix F. While necessarily coarse, empirical results (Section 5.5) confirm its efficacy with minimal impact on performance.

| Method | Text Classification | | | | Image Classification | | | |
|---|---|---|---|---|---|---|---|---|
| | Precision (%) ↑ | Recall (%) ↑ | F1-score (%) ↑ | Time (s) ↓ | Precision (%) ↑ | Recall (%) ↑ | F1-score (%) ↑ | Time (s) ↓ |
| **SAFARI** | **48.3** | **49.3** | **48.5** | 46.37 | 61.4 | 61.1 | 60.5 | 18.64 |
| **SVM** | 47.5 | 47.8 | 47.6 | 91.87 | **63.5** | **62.7** | **62.1** | 69.58 |
| **KNN** | 41.9 | 43.1 | 42.1 | **1.167** | 58.5 | 57.5 | 57.0 | **2.855** |
| **MLP** | 43.4 | 41.6 | 42.1 | 111.5 | 54.7 | 54.4 | 54.3 | 92.55 |
| **RF** | 35.7 | 40.3 | 37.8 | 35.64 | 56.0 | 55.3 | 54.0 | 105.6 |

Table 1: Classification results for text and image modalities. **Bold** and underlined denote the best and second-best scores, respectively. **SAFARI** leads on text classification and ranks second on image classification, balancing accuracy and efficiency.

# 5 Experiments

## 5.1 Datasets and Experiment Environment

**Datasets.** We evaluate **SAFARI** and **SFSes** primarily on five text datasets: **AG-News** (Zhang et al., 2015) (with 4 topic classes: Business, Sci/Tech, Sports, and World), **AAPD** (Yang et al., 2018), **IMDB** (Maas et al., 2011), **Yelp**,[1] and **NewsSpectrum** (Sun et al., 2024). To validate generalizability beyond text, we additionally employ **MIT-States** (Isola et al., 2015) for image embeddings. Further dataset details are available in Appendix A.1.

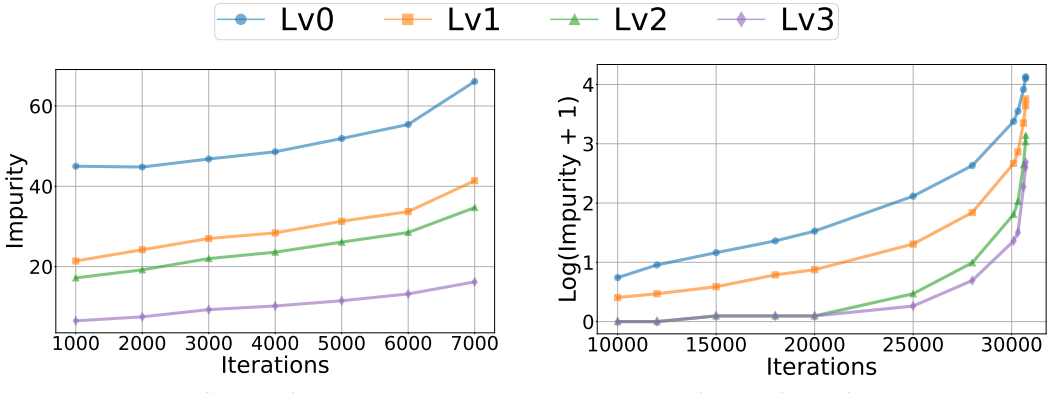
**Experiment Environment.** All methods were implemented in Python 3.8 and evaluated on a Ubuntu 20.04 machine with Intel® Xeon® Platinum 8480C and an NVIDIA H100 GPU.

## 5.2 Hierarchical Semantic Structure Discovery

**Experimental Setup.** We evaluate **SAFARI** and the resulting **SFSes** on **AG-News** for text, and additionally on **MIT-States** for images to test generalizability. Embedding spaces are constructed using BLINK (Wu et al., 2020) for AG-News and CLIP (Radford et al., 2021) for MIT-States. To benchmark hierarchical discovery, we generate 4-level semantic label hierarchies (Lv0 to Lv3) for both datasets using Claude 3.7 Sonnet (Anthropic, 2025), with details provided in Appendix A.2. We assess how well the **SFSes** align with the reference hierarchy using the **impurity** metric:

$$\text{Impurity} = \frac{1}{l}\sum_{i=1}^{l}\left(1 - \frac{1}{|L_i|}\max_{1\le j\le c}|L_i \cap C_j|\right),$$

where $L_i$ is label class $i$; $C_j$ is cluster $j$ used to construct **SFSes**; $l$ and $c$ are the number of label classes and clusters. Lower impurity indicates greater semantic coherence within clusters, with 0 denoting perfect label concentration within clusters. As **SAFARI** progresses, it is expected that impurity grows due to the merging of semantically broader categories, with a consistent ordering across levels: Lv0 (most specific) > Lv1 > Lv2 > Lv3 (most abstract).




(a) AG-News dataset. (b) MIT-States dataset.

Figure 4: Impurity across hierarchical levels.

[1] https://www.yelp.com/dataset

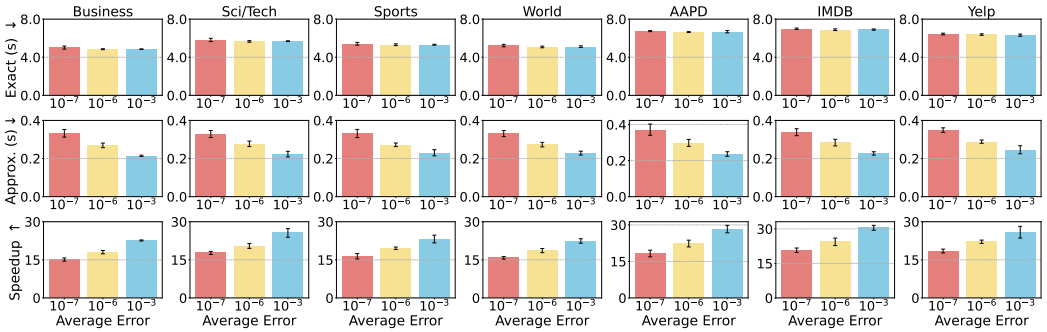



Figure 5: Runtime comparison between exact and approximate Semantic Shift computation across seven topic classes.

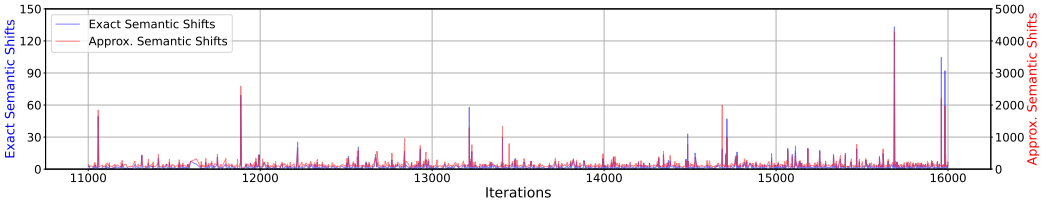



Figure 6: Semantic Shift (exact vs. approximate) on the Sports category in AG-News.

**Results and Analysis.** Figure 4 shows a consistent decrease in impurity from Lv0 to Lv3 across iterations for both modalities. This trend reflects a hierarchical shift from specific to abstract semantics, confirming that **SFSes** capture coherent semantic groupings at multiple granularities.

The consistent patterns across text and image confirm the **generalizability** of **SAFARI**. Without supervision, it uncovers increasingly abstract semantic relationships by tracking Semantic Shifts. By revealing how local neighborhoods compose global hierarchies, **SAFARI** offers a generalizable framework for identifying hierarchical structures in embedding spaces.

### 5.3 Classification

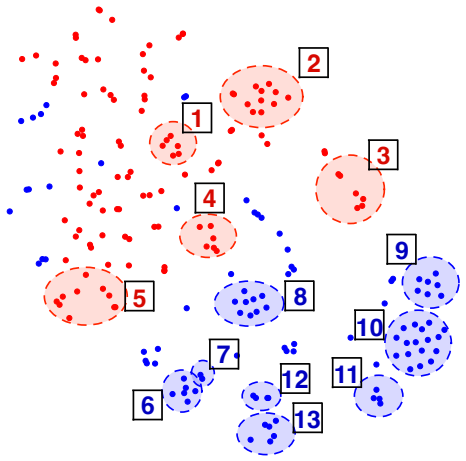
**Experimental Setup.** We measure whether **SFSes** preserve meaningful semantics by testing their performance on classification tasks. For text classification, we use four datasets: **AG-News** (4 topics), **AAPD**, **IMDB**, and **Yelp**, covering seven distinct classes. For image classification, we use **MIT-States**, filtering to 97 object classes with at least 240 samples each. We compare **SAFARI** against four standard classifiers: Support Vector Machine (SVM) (Platt, 1999; Chang & Lin, 2011), K-Nearest Neighbors (KNN) (Cover & Hart, 1967; Fix, 1985), Random Forest (RF) (Breiman, 2001), and Multi-Layer Perceptron (MLP) (He et al., 2015; Hinton, 1990). For classification using **SFSes**, we compute the distance between each test embedding and all identified **SFSes**, assigning the label to the nearest one. Results are presented in Table 1.

**Results and Analysis.** For text classification, **SAFARI** outperforms all baselines while using only around 50% of the computation time. In image classification, it ranks second to SVM but runs 3.7× faster. Overall, **SFSes** offer effective and efficient semantic representations across modalities.

We further evaluate political bias detection on the **NewsSpectrum** dataset (details in Appendix B). Results show that **SAFARI** captures nuanced ideological distinctions where standard classifiers fail, particularly on underrepresented political leanings.

### 5.4 Efficient Semantic Shift Approximation

**Experimental Setup.** To evaluate the efficiency of **SAFARI**'s approximate Semantic Shift computation (Equation 2), we compare it against the full SVD-based method (Equation 1), in terms of runtime. Following the setup in Section 5.3, we sample the top 2,000 entities from each dataset and perform hierarchical clustering. Figure 5 depicts the results, with runtime averaged over 10 independent runs.




| Label | Description |
|---|---|
| 1 | American Eagles men's basketball team |
| 2 | Washington Huskies men's basketball team |
| 3 | Hawaii Rainbow Warriors basket ball team |
| 4 | Connecticut Huskies men's basketball team |
| 5 | Texas A&M Aggies men's basketball team |
| 6 | Iowa State Hawkeyes football team |
| 7 | Iowa State Cyclones football team |
| 8 | Connecticut Huskies men's football team |
| 9 | Alabama Crimson Tide football team |
| 10 | Auburn Tigers football team |
| 11 | Memphis Tigers football team |
| 12 | Penn State Nittany Lions football team |
| 13 | Penn Quakers football team |

Figure 7: SFSes for USA basketball teams and USA football teams.

**Results and Analysis.** As shown in Figure 5, the approximate method achieves a 15∼30× speedup over full SVD across all classes, with average errors below 0.01. This confirms our approximation as a fast, stable alternative to full SVD, making **SAFARI** scalable for large datasets. A detailed analysis of the two Semantic Shift components is provided in Appendix C.

### 5.5 Case Study: Discovering Semantic Hierarchies in AG-News

**Experimental Setup.** To illustrate **SAFARI**'s ability to uncover hierarchical semantics, we conduct a case study on the **Sports** category of the **AG-News** dataset, chosen for its structured, event-driven content. We apply **SAFARI** to the top 2,000 entity embeddings, computing Semantic Shifts at each iteration using both the exact (Equation 1) and approximate (Equation 2) methods. Figure 6 plots Semantic Shift curves between iterations 11,000 and 16,000, with notable evolutions at iterations 11,352 and 15,856. Figure 7 visualizes the U.S. university sports groupings. An additional cross-national example is provided in Appendix E.

**Results and Analysis.** Figure 6 shows that the approximate Semantic Shift curve closely matches the exact one, with a Pearson correlation of **0.92**. **SAFARI**'s sliding-window threshold adapts to non-uniform fluctuations across iterations, enabling the discovery of subtle, meaningful SFSes. In Figure 7, early clusters of specific U.S. university basketball and football teams gradually merge into broader university sports categories, revealing fine-to-coarse semantic organization. These hierarchies are native to the embedding space and should not be expected to follow human ontologies.

## 6 Conclusions

In this paper, we study semantic structure through the evolution of local embedding neighborhoods. **SFSes** represent the semantic context supplied by these neighborhoods, and Semantic Shift characterizes changes in their geometric representations as neighborhoods expand. **SAFARI** tracks these changes during hierarchical clustering to select SFSes at significant transitions and organize semantic fields across scales within the original embedding space. Experiments on five text datasets show that **SFSes** improve classification performance, including political bias detection, while **SAFARI** reveals meaningful semantic hierarchies. Additional image embedding experiments demonstrate applicability across modalities. The Semantic Shift approximation makes discovery computationally efficient. Together, these findings connect local semantic context with broader semantic organization through neighborhood evolution, providing a basis for further embedding analysis and knowledge discovery.

## AI Use Statement

We used GPT solely to improve the grammar and readability of the manuscript. We did not use generative AI to develop the research methodology, implement methods, generate or process data,

formulate mathematical claims or proofs, or interpret results. The authors checked all suggested edits for accuracy and preservation of meaning and accept responsibility for the final manuscript.

## REFERENCES


Anthropic. Claude ai (v3.7). `https://www.anthropic.com`, 2025. Large language model.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations (ICLR)*, 2024.

David A Belsley, Edwin Kuh, and Roy E Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons, 2005.

Leo Breiman. Random Forests. *Machine Learning*, 45:5–32, 2001.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 9912–9924, 2020.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–27, 2011.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pp. 1597–1607, 2020.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286, 2019.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. What is one grain of sand in the desert? Analyzing individual neurons in deep NLP models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 6309–6317, 2019.

Jia Deng, Jonathan Krause, Alexander C Berg, and Li Fei-Fei. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3450–3457, 2012.

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

Ankit Dhall, Anastasia Makarova, Octavian Ganea, Dario Pavllo, Michael Greeff, and Andreas Krause. Hierarchical image classification using entailment cone embeddings. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition Workshops*, pp. 836–837, 2020.

Philipp Dufter and Hinrich Schütze. Analytical methods for interpretable ultradense word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1185–1191, 2019.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65, 2019.

Evelyn Fix. *Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties*, volume 1. USAF School of Aviation Medicine, 1985.

Lu Gan, Diana Nurbakova, Léa Laporte, and Sylvie Calabretto. Enhancing recommendation diversity using determinantal point processes on knowledge graphs. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 2001–2004, 2020.

Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent a new approach to self-supervised learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 21271–21284, 2020.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-Augmented Language Model Pre-Training. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 3929–3938, 2020.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. *arXiv preprint arXiv:0909.4061*, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034, 2015.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738, 2020.

Geoffrey E Hinton. Connectionist learning procedures. In *Machine Learning*, pp. 555–610. Elsevier, 1990.

Kohei Hirata, Daichi Amagata, Sumio Fujita, and Takahiro Hara. Solving diversity-aware maximum inner product search efficiently and effectively. In *Proceedings of the 16th ACM Conference on Recommender Systems (RecSys)*, pp. 198–207, 2022.

Qiang Huang, Yanhao Wang, Yiqun Sun, and Anthony KH Tung. Diversity-aware $k$-maximum inner product search revisited. *arXiv preprint arXiv:2402.13858*, 2024.

Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1383–1391, 2015.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, pp. 4904–4916, 2021.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020.

Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. *Mining of Massive Datasets*. Cambridge University Press, 2020.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 9459–9474, 2020.

Xianming Li and Jing Li. Aoe: Angle-optimized embeddings for semantic textual similarity. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1825–1839, 2024.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations (ICLR)*, 2022.

Tianlin Liu, Lyle Ungar, and Joao Sedoc. Unsupervised post-processing of word vectors via conceptor negation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 6778–6785, 2019.

Jianjie Luo, Yehao Li, Yingwei Pan, Ting Yao, Jianlin Feng, Hongyang Chao, and Tao Mei. Semantic-conditional diffusion networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23359–23368, 2023.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 142–150, 2011.

Carl D Meyer. *Matrix Analysis and Applied Linear Algebra*, volume 71. SIAM, 2000.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS)*, pp. 3111–3119, 2013.

David Mimno and Laure Thompson. The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2873–2878, 2017.

Jiaqi Mu and Pramod Viswanath. All-but-the-Top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations (ICLR)*, 2018.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 2014–2037, 2023.

Eugene A Nida. *A componential analysis of meaning: An introduction to semantic structures*, volume 57. Walter de Gruyter GmbH & Co KG, 2015.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

Sungjoon Park, JinYeong Bak, and Alice Oh. Rotated word vector representations and their interpretability. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 401–411, 2017.

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics (TACL)*, 11:1316–1331, 2023.

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

Adi Simhi and Shaul Markovitch. Interpreting embedding spaces by conceptualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1704–1719, 2023.

Gilbert W Stewart. Perturbation Theory for the Singular Value Decomposition. 1998.

Yiqun Sun, Qiang Huang, Yanhao Wang, and Anthony KH Tung. DiversiNews: Enriching news consumption with relevant yet diverse news articles retrieval. *Proceedings of the VLDB Endowment*, 17(12):4277–4280, 2024.

Lloyd N Trefethen and David Bau. *Numerical Linear Algebra*. SIAM, 2022.

Stephen Ullmann. The principles of semantics. 1957.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European Conference on Computer Vision (ECCV)*, pp. 268–285, 2020.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning (ICML)*, pp. 9929–9939, 2020.

Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6397–6407, 2020.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. SGM: Sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pp. 3915–3926, 2018.

Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. Multimodal transformer with multi-view visual representation for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 30(12):4467–4480, 2019.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 76749–76771, 2023.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS)*, pp. 649–657, 2015.

Yiqian Zhang, Buyu Liu, Jun Bao, Qiang Huang, Min Zhang, and Jun Yu. Learnability matters: Active learning for video captioning. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 37928–37954, 2024.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929, 2016.

# A Experiment Details

## A.1 Dataset Details

We evaluate **SAFARI** on six widely used real-world datasets spanning both text and image modalities, demonstrating its effectiveness and generalizability across domains. For the text modality, we use five diverse datasets:

- **AG-News** (Zhang et al., 2015) consists of over 1 million articles from 2,000+ sources, categorized into four topics: **Business**, **Sci/Tech**, **Sports**, and **World**. This serves as our primary dataset due to its scale and well-defined semantic structures.

- **AAPD** (Yang et al., 2018) contains 55,840 arXiv abstracts from computer science, each labeled with one or more subject areas, supporting multi-label classification tasks.
- **IMDB** (Maas et al., 2011) comprises 50,000 movie reviews split evenly into training and testing sets, designed for binary sentiment classification.[2]
- **Yelp** includes 6.99 million user reviews with 150,000+ business attributes, enabling fine-grained semantic analysis.[3]
- **NewsSpectrum** (Sun et al., 2024) offers 250,000 politically diverse news articles from Reddit. It offers a balanced distribution across the ideological spectrum, making it well-suited for studying abstract semantic phenomena such as political bias.

For the image modality, we use **MIT-States** dataset (Isola et al., 2015), which comprises ∼53,000 images labeled with 245 object classes and 115 attribute states. It is specifically designed to evaluate models on object-attribute compositions and compositional generalization.

### A.2 Hierarchical Label Structures For AG-News and MIT-States

To evaluate semantic coherence across multiple granularities, we construct four-level hierarchies for both datasets:

- **AG-News:** Category → Subcategory → Semantic Group → Entity;
- **MIT-States:** Category → Subcategory → Object → Attribute.

These hierarchies capture increasingly specific relationships, ranging from broad domains to specific entities in **AG-News**, as well as conceptual relationships such as `Materials & Substances` → `Metals` → `Steel` → `Unpainted` in **MIT-States**.

**Prompts for AG-News.** We employ a two-stage prompting strategy using Claude 3.7 Sonnet to construct the AG-News hierarchy. Directly prompting with the full entity list often causes omission due to context length limitations. To mitigate this, we begin with a small subset of entities and prompt Claude to generate a hierarchy with fixed top-level categories (Figure A1). Then, we iteratively expand the hierarchy by combining previous outputs with new subsets of entities (Figure A2), until all entities are processed.

**Prompts for MIT-States.** For MIT-States, we filter out images and categories that lack hierarchical depth, either due to having only one image (e.g., `dog`, `car`) or missing parent categories. The prompt used to construct valid hierarchical relationships is shown in Figure A3.

**Hierarchical Labels for AG-News and MIT-States.** The complete hierarchical label mappings for AG-News and MIT-States are provided in Tables A1 and A2, respectively. These serve as ground truth for evaluating semantic coherence at different levels of abstraction.

### A.3 Detailed Experiment Settings

To comprehensively evaluate **SAFARI** and the effectiveness of **SFSes**, we conduct a series of experiments across diverse datasets and tasks, summarized in Table A3.

**Hierarchical Structure Discovery.** We apply **SAFARI** in a fully unsupervised setting across both text and image modalities. For the **text** modality, we use the AG-News dataset with four categories. To reduce noise from common entities (e.g., `Reuters`) that appear across all categories, we retain only those unique to each class. For the **image** modality, we use MIT-States, comprising 53,000 images across 245 object and 115 attribute classes. Four-level semantic hierarchies (Lv0 to Lv3) are generated for both datasets using Claude 3.7 Sonnet, as detailed in Appendix A.2.

We evaluate the discovered semantic structures using the *impurity* metric that measures label heterogeneity within clusters, ranging from 0 (perfect homogeneity) to higher values (increased mixing). Successful hierarchy discovery is indicated by both increasing diversity values with iterations and a consistent Lv0 $>$ Lv1 $>$ Lv2 $>$ Lv3 ordering that mirrors the progression from specific to abstract concepts.

[2] `https://www.imdb.com/`

[3] `https://www.yelp.com/dataset`

**Prompt 1: Initial Instruction for AG-News**

Here are some entities from news. I want you to label them into 4 levels.

They should have top level (Lv3, Sports, Business, World, Science-Tech).

Then the entities should be grouped and labeled based on their semantics to get Lv2 and Lv1. Lv1 should be the entities themselves.

Avoid any category/label that has only one entity in it.

You should label them directly like a human instead of writing code to do it.

You can put the entities to a Unknown category if you think they do not fit into any of the top 4 categories.

Remember you need to label ALL entities listed.

You need to use the EXACT entity name provided.

Return the result in json format.

Figure A1: Initial prompt to generate a partial AG-News hierarchy with fixed top-level labels.

| Category | Subcategory | Semantic Group |
|---|---|---|
| Sports | Olympic Sports | Teams & Organizations, Events & Competitions, Athletes, Venues |
| | Basketball | Teams & Seasons, Players & Personnel, Venues, Organizations & Events |
| | American Football | Teams & Seasons, Players & Personnel, Venues & Concepts |
| | Baseball | Teams & Seasons, Players & Personnel, Events & Concepts |
| | Other Sports | Golf & Tennis, Racing & Motorsports, Combat Sports, Rugby & Cricket, Horse Racing, Soccer & Football, Swimming & Water Sports, Other Sporting Events, Other Sports Personnel, Winter Sports, Teams & Organizations, Players & Personnel, Venues & Events, Cricket & Rugby, Other Sports Events, Other Sports |
| | Soccer & Football | Teams & Organizations, Players & Personnel, Venues & Events |
| Business | Financial Services | Banking, Investment & Asset Management, Insurance & Risk Management, Consulting & Advisory |
| | Corporations & Industries | Manufacturing & Industrial, Retail & Consumer Goods, Automotive, Transportation & Logistics, Energy & Resources, Technology & Telecommunications |
| | Technology Companies | Software & IT, Security & Cybersecurity, Media & Entertainment, Hardware & Computing, Technology Services |
| | Retail & Consumer Goods | Retail Companies, Food & Beverage, Marketing & Advertising |
| | Corporate Entities | Corporations & Conglomerates, Executives & Entrepreneurs |
| World | Politics & Government | Political Figures, Government Organizations, Political Events & Issues, International Relations, Political Movements |
| | Cultural & Social | Literature & Writers, Social Groups, Media & Entertainment Figures, Arts & Culture, Social Issues |
| | Law & Justice | Legal Cases & Legislation, Legal Professionals, Crime & Legal Issues |
| | Military & Security | Military Conflicts, Organizations, Personnel, Security & Intelligence |
| Science-Tech | Space & Astronomy | Space Exploration, Astronomical Research, Astronomical Objects |
| | Computing & Technology | Software & Development, Hardware & Devices, Internet & Telecom, Digital Media, IT Infrastructure |
| | Medical & Health | Medical Technology, Research, Healthcare Organizations |
| | Environmental Science | Climate & Earth Sciences, Environmental Events |

Table A1: Hierarchical labels for AG-News dataset.

**Classification. SFSes** are constructed using class-labeled training data. Test samples are classified by computing cosine distances to all **SFSes** and assigning the label of the nearest subspace. For text classification, we use all embedding dimensions weighted by singular values; for image classification,

**Prompt 2: Continued Instruction for AG-News**

Here are some entities from news. I want you to label them into 4 levels hierarchy. The entities should be grouped and labeled based on their semantics.

Avoid any category/label that has only one entity in it.

You should label them directly like a human instead of writing code to do it.

You can put the entities to a Unknown category if you think they do not fit into any of the top 4 categories.

Remember you need to label ALL entities listed.

You need to use the EXACT entity name provided.

Return the result in json format.

Existing hierarchy is provided.

But you can create new categories/hierarchy if you think it is necessary.

Figure A2: Second-stage prompt expanding the AG-News hierarchy with additional entity subsets.

**Prompts for MIT-States**

The MIT-States dataset contained only two hierarchical label categories: Attribute and Object.

For example, the object car could have lower-level state variations such as dirty car or new car.

To create a more comprehensive hierarchy, I need you to help me to get higher-level classifications of subcategories and categories.

This allows for broader groupings — for instance, items like beef and apple could be organized under the higher-level category of food.

Here are the object list: ...

Figure A3: Prompt template for constructing MIT-States hierarchies via object-attribute refinement.

we use only the top 5% of dimensions without weighting. This distinction reflects the higher variability and noise (e.g., background features) in image embeddings, which make full-dimensional or weighted comparisons less effective. We report precision, recall, and F1 score, and compare against four standard baselines: Support Vector Machine (**SVM**), K-Nearest Neighbors (**KNN**), Random Forest (**RF**), and Multi-Layer Perceptron (**MLP**).

| Category | Subcategory | Object |
|---|---|---|
| Materials and Substances | Metals | Aluminum, Brass, Bronze, Copper, Metal, Steel |
| | Natural Materials | Clay, Cotton, Fabric, Foam, Paper, Paste, Plastic, Silk, Velvet, Wool |
| | Earth Elements | Concrete, Dirt, Granite, Ground, Mud, Rock, Sand, Stone |
| Food and Consumables | Proteins | Beef, Chicken, Fish, Meat, Salmon, Seafood |
| | Produce | Apple, Fruit, Pear, Tomato, Vegetable, Potato |
| | Prepared Foods | Bread, Cheese, Cookie, Eggs, Pie, Pizza, Soup |
| Built Environment | Structures | Building, Castle, Church, House, Wall |
| | Spaces | Bathroom, Kitchen, Room |
| | Furniture and Fixtures | Cabinet, Chair, Lightbulb, Tile |
| | Transportation Infrastructure | Highway, Road, Street |
| Nature | Bodies of Water | Lake, Pond, Pool |
| | Landforms | Canyon, Valley |
| | Flora | Forest, Plant, Redwood, Tree |
| | Sky Elements | Cloud, Sky |
| | Agricultural | Farm |
| Consumer Goods | Clothing and Accessories | Bracelet, Clothes, Coat, Dress, Necklace, Pants, Ribbon, Ring, Shirt, Shorts |
| | Household Items | Bag, Blade, Bottle, Camera, Carpet, Clock, Glass, Knife, Rope, Toy |

Table A2: Hierarchical labels for the MIT-States dataset.

| Experiment | Datasets | Models / Embeddings | Parameters | Metrics | Baselines |
|---|---|---|---|---|---|
| Hierarchical Structure Discovery in Text | AG-News | BLINK | Dynamic thresholding with sliding window | Impurity | N/A |
| Hierarchical Structure Discovery in Image | MIT-States (53K images) | CLIP | Dynamic thresholding with sliding window | Impurity | N/A |
| Topic Classification | AG-News, AAPD, IMDB, Yelp | BLINK | Top-n entities | Precision, Recall, F1-score | SVM, KNN, RF, MLP |
| Image Classification | MIT-States | CLIP | Top-n dimensions (5%) | Precision, Recall, F1-score | SVM, KNN, RF, MLP |
| Computational Efficiency | AG-News Sampled | BLINK | Dynamic thresholding with sliding window | Runtime, Average error | SAFARI (Full SVD) |
| Political Bias Detection | NewsSpectrum | AnglE | Top-n entities | Runtime, F1-score | SVM, KNN, RF, MLP |
| Component Analysis of Semantic Shift | AG-News Sampled | AnglE | N/A | DIS/DC ratio, Mean, Std, Median | N/A |
| Parameter Study | AG-News Sampled | AnglE | Min window: 50-200, Std mul: 0.5-3.0 | CV, P90/P10, Max/Min ratio | N/A |

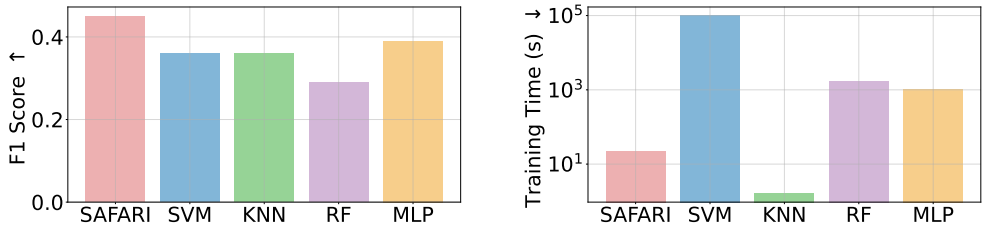
Table A3: Detailed experiment settings.

- **Topic Classification:** We use four text datasets: AG-News, AAPD, IMDB, and Yelp. Entity embeddings are derived using **BLINK** (Wu et al., 2020) for entity linking, followed by TF-IDF-based filtering (Schütze et al., 2008; Leskovec et al., 2020) to retain the top 2,100 entities.
- **Image Classification:** We use the MIT-States dataset with **CLIP** (Radford et al., 2021) to extract image embeddings. We retain 97 object classes with at least 240 samples each to ensure class balance.

For both modalities, embeddings are split into 80% training and 20% testing.

**Computational Efficiency.** We apply **SAFARI** to the sampled AG-News dataset and compare the runtime of our approximate **Semantic Shift** computation against the full SVD-based variant. We report runtime and error rates to quantify the trade-off between efficiency and accuracy.

**Political Bias Detection.** To assess **SAFARI**'s capability in capturing nuanced ideological distinctions beyond surface-level topics, we conduct a political bias detection task using the **NewsSpectrum** dataset. The dataset contains articles classified into five political categories (Left, Lean Left, Center, Lean Right, and Right), with labels derived from media source affiliations rather than individual content analysis. Articles are embedded using the AnglE (Li & Li, 2024) sentence transformer, and the model is evaluated in a supervised classification setup with 80%/20% train-test split. Evaluation metrics include F1 score and training time, and **SAFARI** is again compared with the same four baselines as the classification task.




(a) F1 Scores. (b) Training Time.

Figure B4: Political bias detection results.

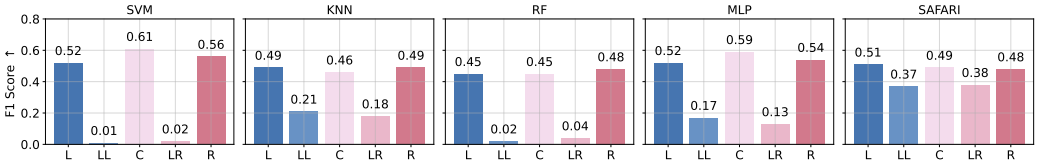



Figure B5: F1 scores across five political bias categories: L (Left), LL (Lean Left), C (Center), LR (Lean Right), and R (Right).

**Component Analysis of Semantic Shift.** To understand the internal mechanisms of **Semantic Shift**, we conduct an experiment to assess the effectiveness of its two components: Dimensional Importance Shift (DIS) and Directional Change (DC) on the sampled **AG-News** dataset. To quantify their relative contributions, we report their mean, median, and standard deviation values.

**Parameter Study.** Finally, we perform a parameter study on the **SAFARI**'s dynamic thresholding mechanism. We vary the minimum window size and test a range of standard deviation multipliers (0.5 to 3.0, with a step of 0.5). To evaluate the uniformity of detected semantic shifts, we report two metrics: Coefficient of Variation (CV) and Max/Min ratio.

## B Political Bias Detection

Beyond conventional classification in Section 5.3, we explore the capability of **SAFARI** in detecting more abstract semantic patterns through political bias detection. Unlike topic classification, where semantic differences are often explicit and content-driven, political bias manifests in subtle linguistic choices and ideological framing that transcend specific topics. This presents a more challenging test for our method: **can SFS identify and represent these nuanced semantics that shape political orientation?**

**Experimental Setup.** We employ **NewsSpectrum** to assess whether SFSes capture abstract semantic patterns that reflect ideological perspectives. Articles are categorized into five bias groups: Left, Lean Left, Center, Lean Right, and Right, though these labels are inherently imprecise as they are assigned based on media sources rather than content. We use AnglE embeddings (Li & Li, 2024), which achieves state-of-the-art performance in the MTEB benchmark (Muennighoff et al., 2023). Since AnglE is a sentence-transformer model, these vectors represent articles rather than entities.

Following the setup in Section 5.3, we compare **SAFARI** against four standard classifiers: **SVM**, **KNN**, **RF**, and **MLP**, using 80% of the data for training and 20% for testing. The results are displayed in Figures B4 and B5.

**Results and Analysis.** The results in Figure B4 highlight that **SAFARI** achieves the best balance between classification performance and computational efficiency for political bias detection. **SAFARI** achieves the highest F1 score (0.45) while maintaining a low training time (21.9 seconds), making it the most practical choice. In comparison, **MLP** reaches a slightly lower F1 score (0.39) but incurs a significantly higher training time (1,033.28 seconds). Other classifiers exhibit varying efficiency-performance trade-offs: **SVM** and **KNN** both have an F1 score of 0.36, yet **SVM** demands extensive training time (27.29 hours), whereas **KNN** is the fastest (1.64 seconds) but offers weaker classification accuracy. **RF** performs the worst, with an F1 score of 0.29 and a substantial training time of 1,684.05

seconds. These results underscore the effectiveness of **SAFARI**, making it scalable and practical for political bias detection tasks.

Figure B5 further validates **SAFARI**'s superior and well-balanced performance in political bias detection compared to standard classifiers. **SAFARI** maintains stable F1 scores across all five categories (ranging from 0.37 to 0.51), whereas standard classifiers exhibit substantial fluctuations and consistently poor performance in lean positions. Notably, **SVM** and **RF** achieve only 0.01–0.04 F1 scores for Lean Left and Lean Right categories, with only **KNN** performing slightly better at 0.21. Furthermore, while some standard classifiers, such as **SVM** and **MLP**, show relatively higher F1 scores for center classification, they struggle significantly with Lean Left and Lean Right classifications. In contrast, **SAFARI** delivers robust and consistent performance across the entire political spectrum. This advantage stems from its unique approach of using **SFSes** to represent different political categories.

Unlike standard classifiers with rigid decision boundaries, **SAFARI** uses subspaces whose basis vectors naturally encode both shared and category-specific semantic patterns. Common linguistic structures or ideological overlaps between categories emerge as shared directions within the subspaces, while category-specific traits are preserved in distinct dimensions. This enables **SAFARI** to recognize when an article aligns semantically with multiple political categories, supporting nuanced, context-aware classification across the political spectrum.

# C Component Analysis of Semantic Shift

**Experimental Setup.** To dissect the inner workings of Semantic Shift, we conduct a controlled experiment that quantifies the individual contributions of its two components: **Dimensional Importance Shift (DIS)** and **Directional Change (DC)**. Our analysis is based on 9,778 valid samples drawn from **SAFARI**'s clustering process, which reveals distinct patterns in their respective impacts. The results are presented in Table C4.

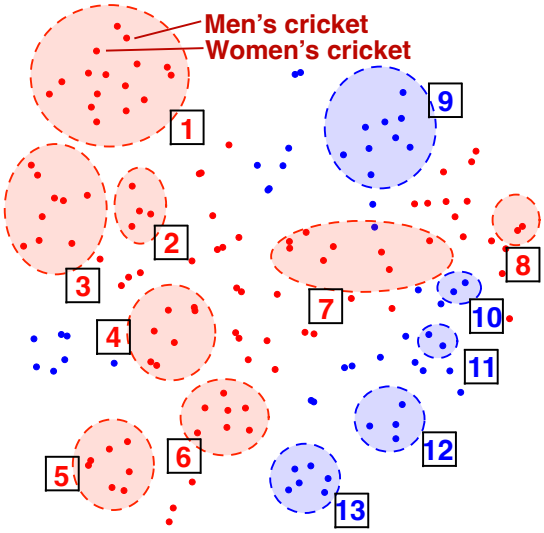
| Metric | DIS | DC |
|---|---|---|
| Mean | 5.60 | 3.19 |
| Median | 3.97 | 2.36 |
| Standard Deviation | 7.41 | 3.25 |
| Log Contribution | 59.3% | 40.7% |

Table C4: Statistics of the contributions of the DIS and DC components.

**Results and Analysis.** Our findings indicate that DIS (5.60) is the dominant factor in stability measurement, contributing approximately 1.8× more than DC (3.19) on average. The right-skewed distribution, evidenced by lower median values compared to means, indicates the presence of significant outliers where DIS is dramatically more influential. While DIS (59.3%) demonstrates greater overall impact, accounting for nearly 60% of the effect, DC (40.7%) remains a substantial contributor at approximately 40%. This confirms that both mechanisms play significant roles in the stability dynamics, though with DIS exerting the stronger influence in most scenarios.

# D Parameter Study about the Dynamic Threshold Mechanism in SAFARI

To robustly detect semantic transitions throughout clustering, **SAFARI** employs a dynamic thresholding mechanism based on a sliding window. Rather than using a fixed window size across all iterations, it applies a recursive divide-and-conquer strategy: the sequence of semantic shifts is iteratively split into smaller segments based on distributional imbalance between halves. This process continues until each segment either reaches a predefined minimum size or exhibits sufficiently balanced distribution. This adaptive strategy prevents extreme Semantic Shifts in later iterations from overshadowing earlier, subtler transitions, ensuring more balanced detection across the entire clustering process.




| Label | Description |
|---|---|
| 1 | Teams of Australia |
| 2 | Teams of Argentina |
| 3 | Teams of Brazil |
| 4 | Teams of Italy |
| 5 | Teams of Japan |
| 6 | Teams of Spain |
| 7 | Teams of Canada |
| 8 | Teams of United States |
| 9 | Teams of Russia |
| 10 | Teams of Finland |
| 11 | Teams of Slovakia |
| 12 | Teams of Lithuania |
| 13 | Teams of Greece |

Figure E6: SFSes for sports teams from non-European and European countries.

We study how two parameters affect the behavior of this dynamic threshold mechanism: (1) the **Standard Deviation Multiplier (SDM)** and (2) the **Minimum Window Size (MWS)** used to define the dynamic threshold. We evaluate the uniformity of detected semantic shifts across different settings to identify configurations that balance sensitivity to meaningful changes with temporal consistency.

**Experimental Setup.** To evaluate the uniformity of Semantic Shifts, we use two complementary metrics: (1) **Coefficient of Variation (CV)**: measures dispersion relative to the mean; lower values imply greater uniformity. (2) **Max/Min Ratio**: captures the full range of detected shift magnitudes; smaller values indicate better balance across iterations. Results are shown in Table D5.

| SDM | MWS | CV ↓ | Max/Min Ratio ↓ |
|---|---|---|---|
| 3.0 | 50-200 | 3.75 | 1,540.6 |
| 2.5 | 50-200 | 4.42 | 1,811.3 |
| 2.0 | 50-200 | 5.10 | 2,193.1 |
| 1.5 | 50-200 | 5.84 | 2,778.6 |
| 1.0 | 50-200 | 6.83 | 3,599.7 |
| 0.5 | 50-200 | 6.78 | 4,384.8 |

Table D5: Performance metrics across standard multipliers.

**Results and Analysis.** The configuration with a standard deviation multiplier of 3.0 consistently achieves the most uniform distribution of Semantic Shifts (CV = 3.75, Max/Min ratio = 1,540.6), regardless of the minimum window size. This setting still ensures that no single phase of the clustering process dominates the detection of **SFSes**. Overall, our parameter study reveals that achieving uniform semantic shift distribution depends primarily on the standard deviation multiplier rather than the minimum window size configuration. The multiplier of 3.0 effectively prevents later iterations from dominating the analysis while maintaining sensitivity to meaningful semantic changes throughout the clustering process.

## E More Analysis on Hierarchical Structure

Figure E6 shows cross-national grouping. Teams initially cluster by country and then merge into regional groups. European teams (blue) consolidate more tightly than non-European teams (red), indicating structure-aware semantic abstraction.

We explore the differences between the hierarchical semantic structures identified by **SAFARI** in embedding spaces and the more intuitive hierarchies found in natural human language. In human language, semantics typically follow a logical hierarchy, progressing from specific, concrete entities

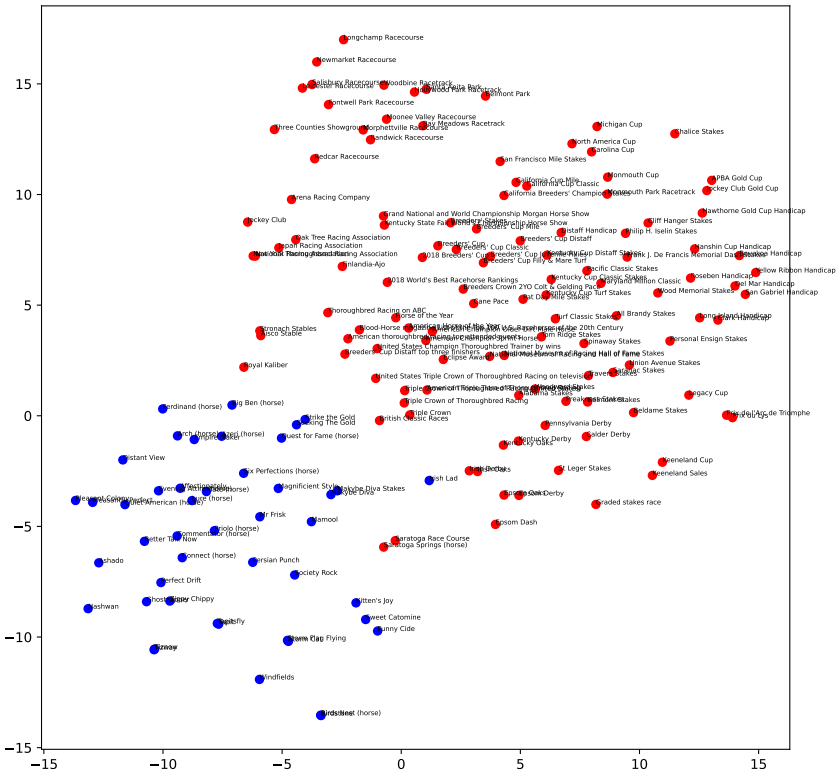



Figure E7: Famous racing horses merged with other horse racing entities.

to more abstract concepts, much like an ontology. However, in embedding spaces, this progression is not always intuitive. The distinction between *specific* and *abstract* depends more on the data and model than on human reasoning, often leading to groupings that diverge from what we would expect based on natural language understanding.

For example, as illustrated in Figure 7, USA basketball teams are first grouped with USA football teams, and later, sports teams from various locations are merged, as shown in Figure E6. This follows a logical hierarchical structure, from more specific categories to broader ones. Yet, as shown in Figure E7 (at iteration 19,790), entities such as horse racing clubs, companies, and events (e.g., `Jockey Club`) are merged with famous racing horses. This merging of horse racing happens thousands of iterations after the merging of football and basketball teams in the USA. Following the ontology-like progression, we would expect more abstract concepts. However, horse racing is not a more abstract concept compared with other sports.

These examples illustrate that the hierarchical structures emerging from embedding spaces are governed by the model's learned representations rather than human-designed logic. While some align intuitively with natural semantic categories, others can be surprising–revealing how models encode relationships that reflect statistical regularities in the data rather than explicit reasoning. This underscores that semantic hierarchies derived from embeddings should not be expected to faithfully mirror human conceptual structures, and should be analyzed with an awareness of the model's inductive biases.

## F Proof of Theorem 4.1

**Proof:** The result follows from Weyl's Theorem (Weyl, 1912), which bounds the change in singular values under additive perturbations. Consider the larger cluster represented by matrix $\boldsymbol{A} \in \mathbb{R}^{m \times d}$. When merging with another cluster, the resulting matrix $\tilde{\boldsymbol{A}}$ can be viewed as a perturbed version of $\boldsymbol{A}$, i.e., $\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{E}$, where $\boldsymbol{E}$ is the perturbed matrix.

**Theorem F1** (Weyl's Theorem (Weyl, 1912))**:** *For any perturbed matrix $\boldsymbol{E}$, the singular values satisfy:* $|\sigma_i(\boldsymbol{A}) - \sigma_i(\tilde{\boldsymbol{A}})| = |\sigma_i(\boldsymbol{A}) - \sigma_i(\boldsymbol{A} + \boldsymbol{E})| \leq \|\boldsymbol{E}\|_2$.

This result implies that the change in any singular value is at most the spectral norm of the perturbed matrix, regardless of its dimension (Stewart, 1998). To apply this, we rewrite $\boldsymbol{A}_{new} = [\boldsymbol{A}_x|\boldsymbol{A}_y] = [\boldsymbol{A}_x|\boldsymbol{O}] + [\boldsymbol{O}|\boldsymbol{A}_y]$, where $\boldsymbol{O}$ is a zero matrix. According to Theorem F1, we have:

$$\begin{aligned} &|\sigma_i([\boldsymbol{A}_x|\boldsymbol{O}]) - \sigma_i(\boldsymbol{A}_{new})| \\ &\quad = |\sigma_i([\boldsymbol{A}_x|\boldsymbol{O}]) - \sigma_i([\boldsymbol{A}_x|\boldsymbol{A}_y])| \\ &\quad \leq \|[\boldsymbol{O}|\boldsymbol{A}_y]\|_2 = \|\boldsymbol{A}_y\|_2 . \end{aligned}$$

As $\sigma_i([\boldsymbol{A}_x|\boldsymbol{O}]) = \sigma_i(\boldsymbol{A}_x)$, Theorem 4.1 is proved. □

**Approximating Directional Change.** Intuitively, $\sum_i d_{sem}(\boldsymbol{v}_i, \tilde{\boldsymbol{v}}_i^*)$ captures the perturbation in basis directions during cluster merging. Following (Meyer, 2000; Belsley et al., 2005), it is expected that a higher sensitivity in $\boldsymbol{A}_x$, typically characterized by its condition number $\kappa(\boldsymbol{A}_x) = \sigma_{max}(\boldsymbol{A}_x)/\sigma_{min}(\boldsymbol{A}_x)$, will result in larger directional changes when $\boldsymbol{A}_x$ and $\boldsymbol{A}_y$ are merged, i.e., $\sum_i d_{sem}(\boldsymbol{v}_i, \tilde{\boldsymbol{v}}_i^*) = \mathcal{O}(\kappa(\boldsymbol{A}_x))$.

Since $\boldsymbol{A}_x$ and $\boldsymbol{A}_y$ represent embedding vectors sampled similarly from the embedding space, the impact of noise on these vectors are similar. Therefore, their minimum singular values can be assumed comparable, i.e, $\sigma_{min}(\boldsymbol{A}_x) \approx \sigma_{min}(\boldsymbol{A}_y) \approx \sigma_{min}(\boldsymbol{A}_{new})$. This allows us to discard the effects of minimum singular values, and the directional change can be approximated by $\sigma_{max}(\boldsymbol{A}_x)$, i.e., $\sigma_{max}(\boldsymbol{A}_x) \approx \mathcal{O}(\kappa(\boldsymbol{A}_x)) \approx \sum_i d_{sem}(\boldsymbol{v}_i, \tilde{\boldsymbol{v}}_i^*)$.