

One Swallow Does Not Make a Summer: Understanding Semantic Structures in Embedding Spaces

Yandong Sun,¹ Qiang Huang,² Ziwei Xu,¹ Yiqun Sun,¹ Yixuan Tang,¹ Anthony K. H. Tung¹

¹School of Computing, National University of Singapore, Singapore

²School of Intelligence Science and Engineering, Harbin Institute of Technology (Shenzhen), Shenzhen, China
 {sun.yandong, ziwei.xu}@u.nus.edu, huangqiang@hit.edu.cn, {sunyq, yixuan, atung}@comp.nus.edu.sg

Abstract

Embedding spaces are fundamental to modern AI, translating raw data into high-dimensional vectors that encode rich semantic relationships. Yet, their internal structures remain opaque, with existing approaches often sacrificing semantic coherence for structural regularity or incurring high computational overhead to improve interpretability. To address these challenges, we introduce the **Semantic Field Subspace (SFS)**, a geometry-preserving, context-aware representation that captures local semantic neighborhoods within the embedding space. We also propose **SAFARI (SemAntic Field subspAce deteRmInation)**, an unsupervised, modality-agnostic algorithm that uncovers hierarchical semantic structures using a novel metric called Semantic Shift, which quantifies how semantics evolve as SFSes evolve. To ensure scalability, we develop an efficient approximation of Semantic Shift that replaces costly SVD computations, achieving a $15\sim30\times$ speedup with average errors below 0.01. Extensive evaluations across six real-world text and image datasets show that SFSes outperform standard classifiers not only in classification but also in nuanced tasks such as political bias detection, while SAFARI consistently reveals interpretable and generalizable semantic hierarchies. This work presents a unified framework for structuring, analyzing, and scaling semantic understanding in embedding spaces.

1 Introduction

Embedding spaces are foundational to modern AI systems, converting unstructured inputs—such as text, images, and time series—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 across modalities. 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.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

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 and Viswanath 2018; Liu, Ungar, and Sedoc 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, Bak, and Oh 2017; Dufter and 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 and 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.

Despite substantial progress, embedding spaces remain intrinsically opaque due to three fundamental challenges:

- **Abstract Semantics:** Embeddings inhabit high dimensional spaces where complex, abstract relationships emerge, defying straightforward interpretation. Many existing methods enhance interoperability by restructuring embeddings, but in doing so, they often distort the native geometry, thereby undermining their practical utility.
- **Lack of Explicit Structure:** While semantics are inherently structured, real-world embeddings often exhibit diffuse and irregular patterns. Existing methods either overlook this latent structure or impose rigid taxonomies, limiting flexibility across tasks and domains.
- **Limited Modality Generalization:** Semantic meaning transcends text to include visual, auditory, and other modalities. However, most techniques are modality-specific and lack a unified framework for revealing semantic structures across diverse embedding spaces.

This paper investigates semantic structures directly within native embedding spaces, without re-embedding, restructuring, or imposing external constraints that alter their original geometry. To address the abstract and opaque nature of high-dimensional embeddings, we introduce a new semantic representation, **Semantic Fields Subspaces** (**SFSes**), along with **SAFARI** (SemAntic Field subspAce deteRmInation), an unsupervised, modality-agnostic algorithm that discovers and organizes semantic structures hierarchically. Our key contributions are as follows:

- **Interpretable Semantic Representation:** We introduce **SFSes**, a context-aware, geometry-preserving representation that captures semantic meaning through local neighborhoods, offering interpretability without distorting the embedding space.
- **Unsupervised Hierarchical Structure Discovery:** We propose **SAFARI**, which leverages a novel *Semantic Shift* metric to uncover hierarchical structures. A scalable approximation of Semantic Shift enables **SAFARI** to process large datasets with minimal accuracy loss.
- **Modality-Agnostic Generalization:** **SFSes** and **SAFARI** are inherently modality-agnostic, uncovering hierarchical semantic structures across both text and image modalities without supervision or external ontologies.

We validate our framework on six real-world datasets across text and image modalities. **SAFARI** uncovers how local neighborhoods form meaningful global hierarchies, while **SFSes** outperform standard classifiers on text classification, deliver competitive image classification with far lower computational cost, and capture subtle semantics (e.g., political bias) often missed by conventional methods. Our Semantic Shift approximation achieves a $15\sim30\times$ speedup over full SVD, with average errors below 0.01, ensuring both efficiency and accuracy. Together, **SFSes** and **SAFARI** form a unified, interpretable, and scalable framework for semantic understanding within embedding spaces.

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 focused on the geometric properties of embeddings, particularly in NLP. Mu and Viswanath (2018) mitigated anisotropy by removing dominant principal components, while Liu, Ungar, and Sedoc (2019) stabilized embedding distributions by suppressing high-variance dimensions. Ethayarajah (2019) found that contextualized embeddings often cluster in narrow cones, limiting expressiveness. To improve interpretability, rotation-based alignment (Park, Bak, and Oh 2017; Dufter and 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 and Isola (2020) analyzed such effects in contrastive learning, while feature visualization methods (Olah, Mordvintsev, and Schubert 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 and 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, and modality-agnostic 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 are modality-agnostic. Let $h : \mathcal{T} \rightarrow \mathcal{E}$ 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 $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ is defined as $d_{sem}(\mathbf{u}, \mathbf{v}) = 1 - \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\| \|\mathbf{v}\|}$.*

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.*

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. \triangle

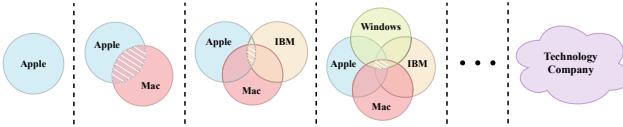


Figure 1: Contextual interpretation of Apple: Meaning refines as more related terms are introduced.

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 (\mathbf{v}) from those representing real-world terms (\mathbf{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 $\mathbf{v}_t \in \mathcal{F}$ such that for all $\mathbf{u}_t \in \mathcal{F}$, we have $d_{sem}(\mathbf{u}_t, \mathbf{v}_t) < \epsilon$.*

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 and 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.

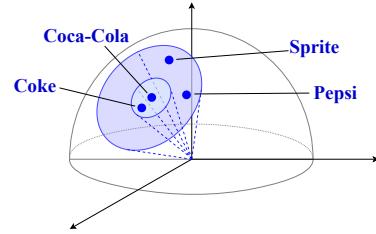


Figure 2: Illustration of Semantic Field exploration.

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

We compute the basis of $\mathbb{S}_{\mathcal{F}}$ via SVD on matrix $M_{\mathcal{F}} = [\mathbf{v}_1, \dots, \mathbf{v}_n]$, i.e., $M_{\mathcal{F}} = \mathbf{U}\Sigma\mathbf{V}^\top$. This yields a continuous, low-rank representation that captures semantic structure while remaining invariant to redundancy.

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.*

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:*

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

where $\Delta \sigma_i = |\sigma_i - \tilde{\sigma}_i|$ captures the **dimensional importance shift** in singular values $\sigma_i \in \Sigma_x$ and $\tilde{\sigma}_i \in \Sigma_{new}$;

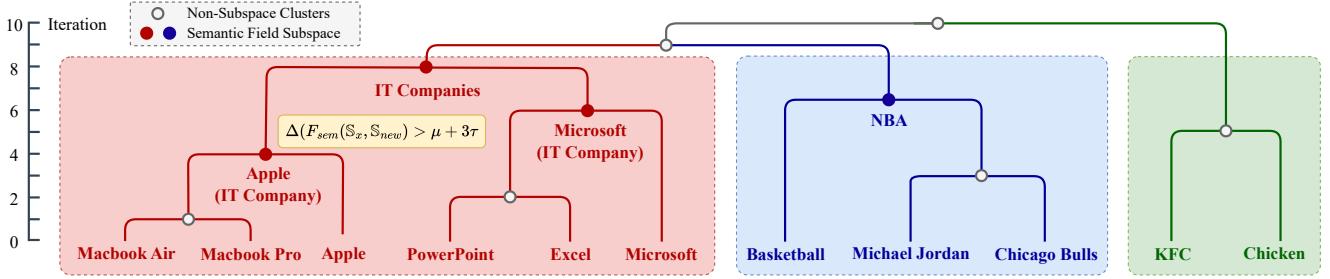


Figure 3: Toy example illustrating **SAFARI**'s hierarchical clustering process.

$d_{sem}(\mathbf{v}_i, \tilde{\mathbf{v}}_i^*)$ captures **directional change** between basis vectors $\mathbf{v}_i \in V_x^\top$ and their nearest counterparts $\tilde{\mathbf{v}}_i^* \in V_{new}^\top$.

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 than $\mathbb{S}_{\mathcal{F}_x}$, 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 Ω , and maintains a set Φ 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 Ω .
- **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 μ and standard deviation τ . 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 Φ 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)

Algorithm 1: SAFARI

```

Input: Embedding set  $\mathcal{E} \subset \mathbb{R}^d$ , window size  $w$ ;
Output: Set  $\Phi$  of Semantic Field Subspaces (SFSes);
1  $\Omega \leftarrow$  Initialize each  $\mathbf{v}_t \in \mathcal{E}$  as its own cluster;
2  $\mu \leftarrow 0$ ;  $\tau \leftarrow 0$ ;  $\Phi \leftarrow \emptyset$ ;
3 while  $|\Omega| > 1$  do
    ▷ Step 1: Cluster Merging
    4  $\{\mathcal{C}_x, \mathcal{C}_y\} \leftarrow \arg \min_{\mathcal{C}_i, \mathcal{C}_j \in \Omega} d_{sem}(\mathcal{C}_i, \mathcal{C}_j)$ ;
    5  $\mathcal{C}_{new} \leftarrow \mathcal{C}_x \cup \mathcal{C}_y$ ;
    6  $\Omega \leftarrow \Omega \setminus \{\mathcal{C}_{new}\} \cup \{\mathcal{C}_x, \mathcal{C}_y\}$ ;
    ▷ Step 2: SFS Delineation
    7  $\mathcal{C}_x \leftarrow |\mathcal{C}_x| > |\mathcal{C}_y| ? \mathcal{C}_x : \mathcal{C}_y$ ;
    8  $\mathbb{S}_x, \mathbb{S}_{new} \leftarrow \text{span}(\mathcal{C}_x), \text{span}(\mathcal{C}_{new})$ ;
    9 Compute  $\Delta F_{sem}(\mathbb{S}_x, \mathbb{S}_{new})$  using Eq. 1;
    10 if  $\Delta F_{sem}(\mathbb{S}_x, \mathbb{S}_{new}) > \mu + 3\tau$  then
        11    $\Phi \leftarrow \Phi \cup \{\mathbb{S}_{new}\}$ ;
    12 Update  $\mu$  and  $\tau$  using the last  $w$  values of  $\Delta F_{sem}$ ;
13 return  $\Phi$ ;

```

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 prevents the merge between IT Companies and NBA, preserving semantic boundaries. \triangle

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, Martinsson, and Tropp 2009; Trefethen and Bau 2022). This becomes a major bottleneck in large-scale applications. To improve scalability, we propose a practical approximation. Let A_x

and \mathbf{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:

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

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

Theoretical Justification. Let $\mathbf{A}_{new} = [\mathbf{A}_x | \mathbf{A}_y]$ be the matrix representing the newly merged cluster \mathcal{C}_{new} . We now justify the approximation by establishing two key theoretical results:

- An upper bound on the dimensional importance shift in singular values;
- A connection between directional change and the largest singular value of \mathbf{A}_x .

Bounding Dimensional Importance Shift. We begin with the following result:

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

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

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 $\mathbf{A} \in \mathbb{R}^{m \times d}$. When merging with another cluster, the resulting matrix $\tilde{\mathbf{A}}$ can be viewed as a perturbed version of \mathbf{A} , i.e., $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{E}$, where \mathbf{E} is the perturbed matrix.

Theorem 4.2 (Weyl's Theorem (Weyl 1912)): *For any perturbed matrix \mathbf{E} , the singular values satisfy: $|\sigma_i(\mathbf{A}) - \sigma_i(\tilde{\mathbf{A}})| = |\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{A} + \mathbf{E})| \leq \|\mathbf{E}\|_2$.*

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 $\mathbf{A}_{new} = [\mathbf{A}_x | \mathbf{A}_y] = [\mathbf{A}_x | \mathbf{O}] + [\mathbf{O} | \mathbf{A}_y]$, where \mathbf{O} is a zero matrix. According to Theorem 4.2, we have:

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

Since $\sigma_i([\mathbf{A}_x | \mathbf{O}]) = \sigma_i(\mathbf{A}_x)$, Theorem 4.1 is proved. \square

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

Since \mathbf{A}_x and \mathbf{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, $\sigma_{min}(\mathbf{A}_x) \approx$

$\sigma_{min}(\mathbf{A}_y) \approx \sigma_{min}(\mathbf{A}_{new})$. This allows us to discard the effects of minimum singular values, and the directional change can be approximated by $\sigma_{max}(\mathbf{A}_x)$, i.e., $\sigma_{max}(\mathbf{A}_x) \approx \mathcal{O}(\kappa(\mathbf{A}_x)) \approx \sum_i d_{sem}(\mathbf{v}_i, \tilde{\mathbf{v}}_i^*)$.

While this approximation is necessarily coarse, it offers an efficient alternative for real-time clustering. Empirical results (Section 5.5) confirm its efficacy with minimal impact on performance.

5 Experiments

5.1 Datasets and Experiment Environment

Datasets. We evaluate **SAFARI** and **SFSes** on six public datasets across text and image modalities. For the text modality, we use five diverse datasets: **AG-News** (Zhang, Zhao, and LeCun 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). For the image modality, we employ **MIT-States** (Isola, Lim, and Adelson 2015), which enables the evaluation of object-attribute composition in visual 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 two modality-diverse datasets: **AG-News** for text and **MIT-States** for images. 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:

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

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).

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 **modality-agnostic** nature of **SAFARI**. Without supervision, it uncovers increasingly abstract semantic relationships by tracking Semantic Shifts. By revealing how local

¹<https://www.yelp.com/dataset>

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	<u>47.5</u>	47.8	<u>47.6</u>	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.

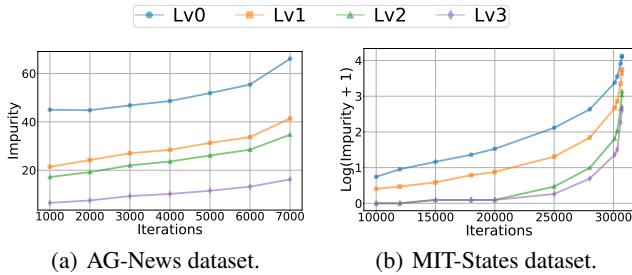


Figure 4: Impurity across hierarchical levels.

neighborhoods compose global hierarchies, **SAFARI** offers a robust and generalizable framework for identifying hierarchical structures in embedding spaces, advancing our understanding of their inherent semantic organization.

5.3 Classification Across Modalities

Experimental Setup. We measure whether **SFSes** preserve meaningful semantics by testing their performance on classification tasks across text and image modalities. 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 and Lin 2011), K-Nearest Neighbors (KNN) (Cover and 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, surpassing SVM, the second-best, while using only around 50% of the computation time. In image classification, it ranks second to SVM in accuracy but runs $3.7 \times$ faster, demonstrating a strong accuracy-efficiency trade-off. Overall, **SAFARI** delivers competitive performance across modalities with notable computational savings, confirming that **SFSes** offer effective and efficient semantic representations.

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

larly on underrepresented political leanings, highlighting its robustness in modeling subtle, real-world semantics beyond surface-level topics.

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.

Results and Analysis. As shown in Figure 5, the approximate method achieves a $15\sim30\times$ speedup over full SVD across all classes, with consistently low variance (as indicated by the error bars). Despite the substantial acceleration, the average error between exact and approximate Semantic Shifts remains below 0.01, within the 10^{-3} scale, ensuring strong accuracy-efficiency trade-offs. These results confirm that our approximation is a fast, stable, and reliable alternative to full SVD, making **SAFARI** scalable for large datasets. A detailed analysis of the two Semantic Shift components, i.e., Dimensional Importance Shift and Directional Change, 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. Figures 7 and 8 visualize the corresponding hierarchical groupings.

Results and Analysis. Figure 6 shows that the approximate Semantic Shift curve closely mirrors the exact one, achieving a Pearson correlation of **0.92**. This strong correlation confirms the reliability of our efficient approximate method introduced in Section 4.3. Moreover, **SAFARI**'s sliding-window-based dynamic thresholding effectively adapts to the non-uniform fluctuation of shifts across iterations, enable the discovery of subtle yet meaningful **SFSes** that would be missed by static thresholds. Further parameter

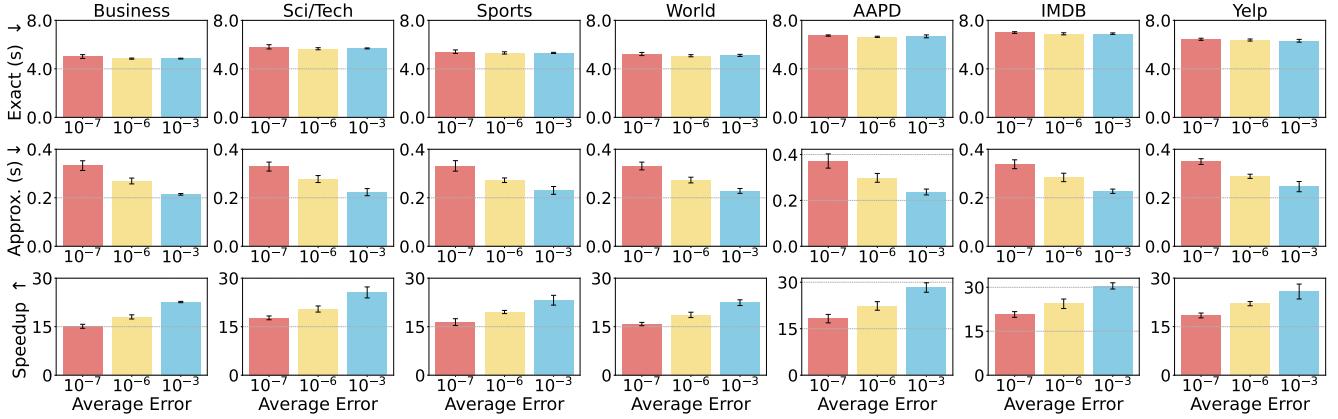


Figure 5: Runtime comparison between exact and approximate Semantic Shift computation across seven topic classes.

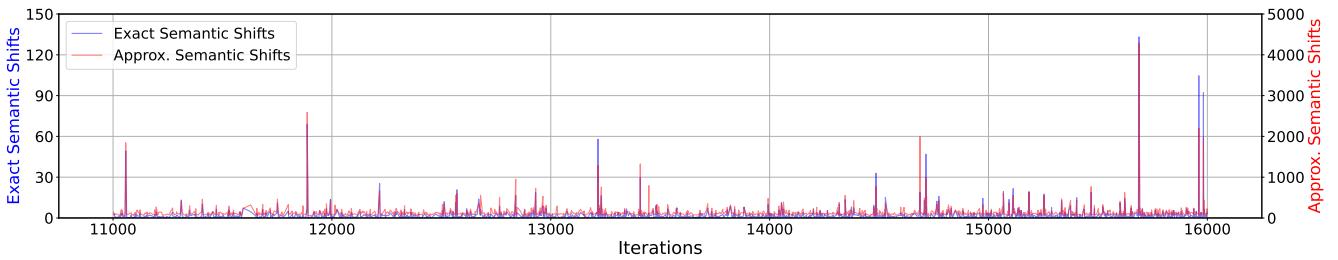


Figure 6: Semantic Shift (exact vs. approximate) on the Sports category in AG-News.

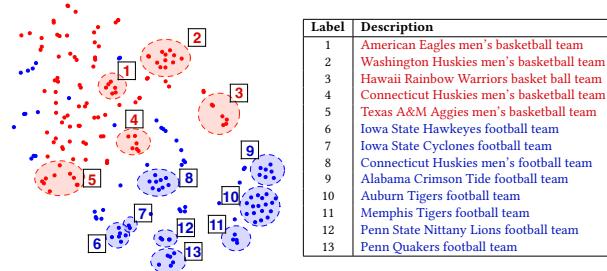


Figure 7: SFSe for USA basketball teams and USA football teams.

study about the dynamic threshold mechanism in **SAFARI** is detailed in Appendix D.

SAFARI also reveals hierarchical semantic relationships with fine-to-coarse granularity. In Figure 7, early clusters capture specific U.S. university basketball and football teams, which gradually merge into broader categories like university sports teams. Figure 8 illustrates cross-national grouping: teams initially cluster by country, then merge into regional groupings. Notably, European teams (blue) consolidate more tightly than non-European teams (red), indicating structure-aware semantic abstraction. These results showcase **SAFARI**'s ability to track evolving semantic organization, understanding latent hierarchies in embedding spaces. Additional analysis is provided in Appendix E.

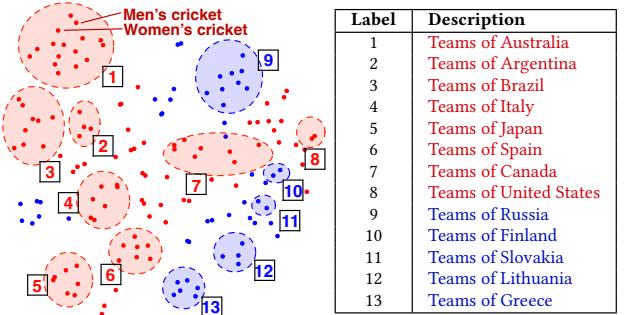


Figure 8: SFSe for sports teams from non-European and European countries.

6 Conclusions

In this paper, we tackle the fundamental challenge of understanding the abstract and intricate structure of embedding spaces. We introduce **SFSe**s as a structured representation that explicitly links embedding spaces to their underlying semantics. Leveraging hierarchical clustering and the concept of Semantic Shift, we develop **SAFARI**, an effective algorithm that uncovers hierarchical semantic structures while maintaining computational scalability through an efficient approximation of Semantic Shift. Through comprehensive experiments on six real-world datasets spanning text and image modalities, we show that **SFSe**s improve per-

formance on both standard classification tasks and subtle semantic challenges like political bias detection. **SAFARI** consistently reveals meaningful, modality-agnostic semantic hierarchies with minimal computational overhead. By bridging the gap between geometric embedding representations and their underlying semantics, this work opens new avenues for future research, like semantic-aware embedding analysis and knowledge discovery.

References

- Anthropic. 2025. Claude AI (v3.7). <https://www.anthropic.com>. Large language model.
- Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; and Hajishirzi, H. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *International Conference on Learning Representations (ICLR)*.
- Belsley, D. A.; Kuh, E.; and Welsch, R. E. 2005. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons.
- Breiman, L. 2001. Random Forests. *Machine Learning*, 45: 5–32.
- Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, 9912–9924.
- Chang, C.-C.; and Lin, C.-J. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3): 1–27.
- Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 1597–1607.
- Clark, K.; Khandelwal, U.; Levy, O.; and Manning, C. D. 2019. 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*, 276–286.
- Cover, T.; and Hart, P. 1967. Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, 13(1): 21–27.
- Dalvi, F.; Durrani, N.; Sajjad, H.; Belinkov, Y.; Bau, A.; and Glass, J. 2019. 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)*, 6309–6317.
- Deng, J.; Krause, J.; Berg, A. C.; and Fei-Fei, L. 2012. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3450–3457.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. 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)*, 4171–4186.
- Dhall, A.; Makarova, A.; Ganea, O.; Pavllo, D.; Greeff, M.; and Krause, A. 2020. Hierarchical image classification using entailment cone embeddings. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition Workshops*, 836–837.
- Dufter, P.; and Schütze, H. 2019. 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)*, 1185–1191.
- Ethayarajh, K. 2019. 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)*, 55–65.
- Fix, E. 1985. *Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties*, volume 1. USAF School of Aviation Medicine.
- Gan, L.; Nurbakova, D.; Laporte, L.; and Calabretto, S. 2020. 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)*, 2001–2004.
- Goh, G.; Cammarata, N.; Voss, C.; Carter, S.; Petrov, M.; Schubert, L.; Radford, A.; and Olah, C. 2021. Multimodal Neurons in Artificial Neural Networks. *Distill*, 6(3): e30.
- Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. A.; Guo, Z. D.; Azar, M. G.; et al. 2020. Bootstrap your own latent a new approach to self-supervised learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, 21271–21284.
- Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M.-W. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 3929–3938.
- Halko, N.; Martinsson, P.-G.; and Tropp, J. A. 2009. Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. *arXiv preprint arXiv:0909.4061*.
- He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 9729–9738.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1026–1034.
- Hinton, G. E. 1990. Connectionist learning procedures. In *Machine Learning*, 555–610. Elsevier.

- Hirata, K.; Amagata, D.; Fujita, S.; and Hara, T. 2022. Solving Diversity-Aware Maximum Inner Product Search Efficiently and Effectively. In *Proceedings of the 16th ACM Conference on Recommender Systems (RecSys)*, 198–207.
- Huang, Q.; Wang, Y.; Sun, Y.; and Tung, A. K. 2024. Diversity-Aware k -Maximum Inner Product Search Revisited. *arXiv preprint arXiv:2402.13858*.
- Isola, P.; Lim, J. J.; and Adelson, E. H. 2015. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 1383–1391.
- Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, 4904–4916.
- Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781.
- Leskovec, J.; Rajaraman, A.; and Ullman, J. D. 2020. *Mining of Massive Datasets*. Cambridge University Press.
- Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktaschel, T.; et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, 9459–9474.
- Li, X.; and Li, J. 2024. AoE: Angle-optimized embeddings for semantic textual similarity. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 1825–1839.
- Li, Y.; Liang, F.; Zhao, L.; Cui, Y.; Ouyang, W.; Shao, J.; Yu, F.; and Yan, J. 2022. Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm. In *International Conference on Learning Representations (ICLR)*.
- Liu, T.; Ungar, L.; and Sedoc, J. 2019. Unsupervised post-processing of word vectors via concept negation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 6778–6785.
- Luo, J.; Li, Y.; Pan, Y.; Yao, T.; Feng, J.; Chao, H.; and Mei, T. 2023. Semantic-conditional diffusion networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23359–23368.
- Maas, A.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, 142–150.
- Meyer, C. D. 2000. *Matrix Analysis and Applied Linear Algebra*, volume 71. SIAM.
- Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS)*, 3111–3119.
- Mimno, D.; and Thompson, L. 2017. The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2873–2878.
- Mu, J.; and Viswanath, P. 2018. All-but-the-Top: Simple and Effective Postprocessing for Word Representations. In *International Conference on Learning Representations (ICLR)*.
- Muennighoff, N.; Tazi, N.; Magne, L.; and Reimers, N. 2023. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2014–2037.
- Nida, E. A. 2015. *A componential analysis of meaning: An introduction to semantic structures*, volume 57. Walter de Gruyter GmbH & Co KG.
- Olah, C.; Mordvintsev, A.; and Schubert, L. 2017. Feature Visualization. *Distill*, 2(11): e7.
- Park, S.; Bak, J.; and Oh, A. 2017. Rotated word vector representations and their interpretability. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 401–411.
- Platt, J. 1999. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers*, 10(3): 61–74.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 8748–8763.
- Ram, O.; Levine, Y.; Dalmedigos, I.; Muhlgay, D.; Shashua, A.; Leyton-Brown, K.; and Shoham, Y. 2023. In-Context Retrieval-Augmented Language Models. *Transactions of the Association for Computational Linguistics (TACL)*, 11: 1316–1331.
- Schütze, H.; Manning, C. D.; and Raghavan, P. 2008. *Introduction to Information Retrieval*. Cambridge University Press.
- Simhi, A.; and Markovitch, S. 2023. Interpreting Embedding Spaces by Conceptualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1704–1719.
- Stewart, G. W. 1998. Perturbation Theory for the Singular Value Decomposition.
- Sun, Y.; Huang, Q.; Wang, Y.; and Tung, A. K. 2024. DiversiNews: Enriching News Consumption with Relevant Yet Diverse News Articles Retrieval. *Proceedings of the VLDB Endowment*, 17(12): 4277–4280.
- Trefethen, L. N.; and Bau, D. 2022. *Numerical Linear Algebra*. SIAM.
- Ullmann, S. 1957. The Principles of Semantics.
- Van Gansbeke, W.; Vandenhende, S.; Georgoulis, S.; Proesmans, M.; and Van Gool, L. 2020. SCAN: Learning to Classify Images Without Labels. In *European Conference on Computer Vision (ECCV)*, 268–285.

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

Weyl, H. 1912. Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Mathematische Annalen*, 71(4): 441–479.

Wu, L.; Petroni, F.; Josifoski, M.; Riedel, S.; and Zettlemoyer, L. 2020. Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6397–6407.

Yang, P.; Sun, X.; Li, W.; Ma, S.; Wu, W.; and Wang, H. 2018. SGM: Sequence Generation Model for Multi-label Classification. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, 3915–3926.

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

Yu, S.; Cho, J.; Yadav, P.; and Bansal, M. 2023. Self-chained image-language model for video localization and question answering. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*, 76749–76771.

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

Zhang, Y.; Liu, B.; Bao, J.; Huang, Q.; Zhang, M.; and Yu, J. 2024. Learnability Matters: Active Learning for Video Captioning. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 37928–37954.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2921–2929.

Reproducibility Checklist

Instructions for Authors:

This document outlines key aspects for assessing reproducibility. Please provide your input by editing this .tex file directly.

For each question (that applies), replace the “Type your response here” text with your answer.

Example: If a question appears as

```
\question{Proofs of all novel claims  
are included} {(yes/partial/no)}  
Type your response here
```

you would change it to:

```
\question{Proofs of all novel claims  
are included} {(yes/partial/no)}  
yes
```

Please make sure to:

- Replace ONLY the “Type your response here” text and nothing else.
- Use one of the options listed for that question (e.g., yes, no, partial, or NA).
- **Not** modify any other part of the \question command or any other lines in this document.

You can \input this .tex file right before \end{document} of your main file or compile it as a stand-alone document. Check the instructions on your conference’s website to see if you will be asked to provide this checklist with your paper or separately.

1. General Paper Structure.

- 1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) **yes**
- 1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) **yes**
- 1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) **yes**

2. Theoretical Contributions.

- 2.1. Does this paper make theoretical contributions? (yes/no) **yes**
If yes, please address the following points:
 - 2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) **yes**
 - 2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) **yes**
 - 2.4. Proofs of all novel claims are included (yes/partial/no) **yes**
 - 2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) **yes**
 - 2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) **yes**
 - 2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) **yes**
 - 2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) **yes**

3. Dataset Usage.

3.1. Does this paper rely on one or more datasets? (yes/no)

yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) **yes**

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) **NA**

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) **NA**

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) **yes**

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) **yes**

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisfying (yes/partial/no/NA) **NA**

4. Computational Experiments.

4.1. Does this paper include computational experiments? (yes/no) **yes**

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) **yes**

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) **yes**

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) **yes**

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) **yes**

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) **yes**

4.7. If an algorithm depends on randomness, then the

method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) **NA**

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) **yes**

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) **yes**

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) **yes**

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) **no**

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) **no**

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) **yes**

Experiment Details

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, Zhao, and LeCun 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

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

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

- **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, Lim, and Adelson 2015), which comprises $\sim 53,000$ images labeled with 245 object classes and 115 attribute states. It is specifically designed to evaluate models on object-attribute compositions and compositional generalization.

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;

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

- **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 1). Then, we iteratively expand the hierarchy by combining previous outputs with new subsets of entities (Figure 2), 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 3.

Hierarchical Labels for AG-News and MIT-States. The complete hierarchical label mappings for AG-News and MIT-States are provided in Tables 1 and 2, respectively. These serve as ground truth for evaluating semantic coherence at different levels of abstraction.

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 3.

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 4.

Table 1: Hierarchical labels for AG-News dataset.

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
Business	Soccer & Football	Teams & Organizations, Players & Personnel, Venues & Events
	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
World	Corporate Entities	Corporations & Conglomerates, Executives & Entrepreneurs
	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

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 $\text{Lv0} > \text{Lv1} > \text{Lv2} > \text{Lv3}$ ordering that mirrors the progression from specific to abstract concepts.

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

- **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, Manning, and Raghavan 2008; Leskovec, Rajaraman,

and Ullman 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 and 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

Table 2: Hierarchical labels for the MIT-States dataset.

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
Consumer Goods	Agricultural	Farm
	Clothing and Accessories	Bracelet, Clothes, Coat, Dress, Necklace, Pants, Ribbon, Ring, Shirt, Shorts
Consumer Goods	Household Items	Bag, Blade, Bottle, Camera, Carpet, Clock, Glass, Knife, Rope, Toy

the classification task.

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.

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 and Li 2024), which achieves state-of-the-art performance in the MTEB benchmark (Muenninghoff et al. 2023). Since AnglE is a sentence-transformer model, these vectors represent articles rather than entities.

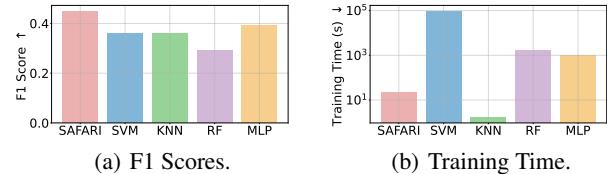


Figure 4: Political bias detection results.

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 4 and 5.

Results and Analysis. The results in Figure 4 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 5 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. No-

Table 3: Detailed experiment settings.

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 images (53K)	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	AngleIE	Top-n entities	Runtime, F1-score	SVM, KNN, RF, MLP
Component Analysis of Semantic Shift	AG-News Sampled	AngleIE	N/A	DIS/DC ratio, Mean, Std, Median	N/A
Parameter Study	AG-News Sampled	AngleIE	Min window: 50-200, Std mul: 0.5-3.0	CV, P90/P10, Max/Min ratio	N/A

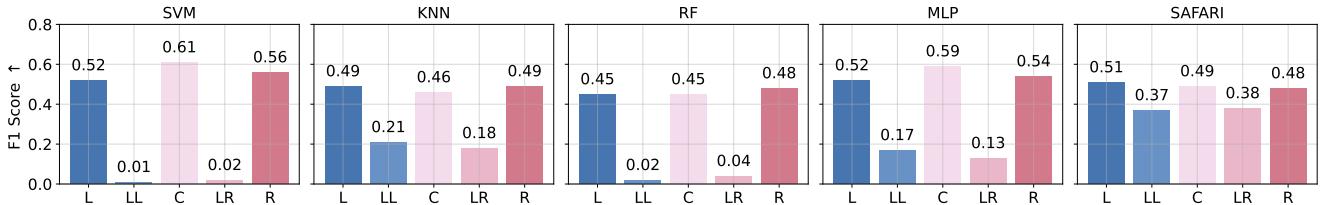


Figure 5: F1 scores across five political bias categories: L (Left), LL (Lean Left), C (Center), LR (Lean Right), and R (Right).

ably, **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.

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 compo-

Table 4: Statistics of the contributions of the DIS and DC components.

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%

nents: **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 4.

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

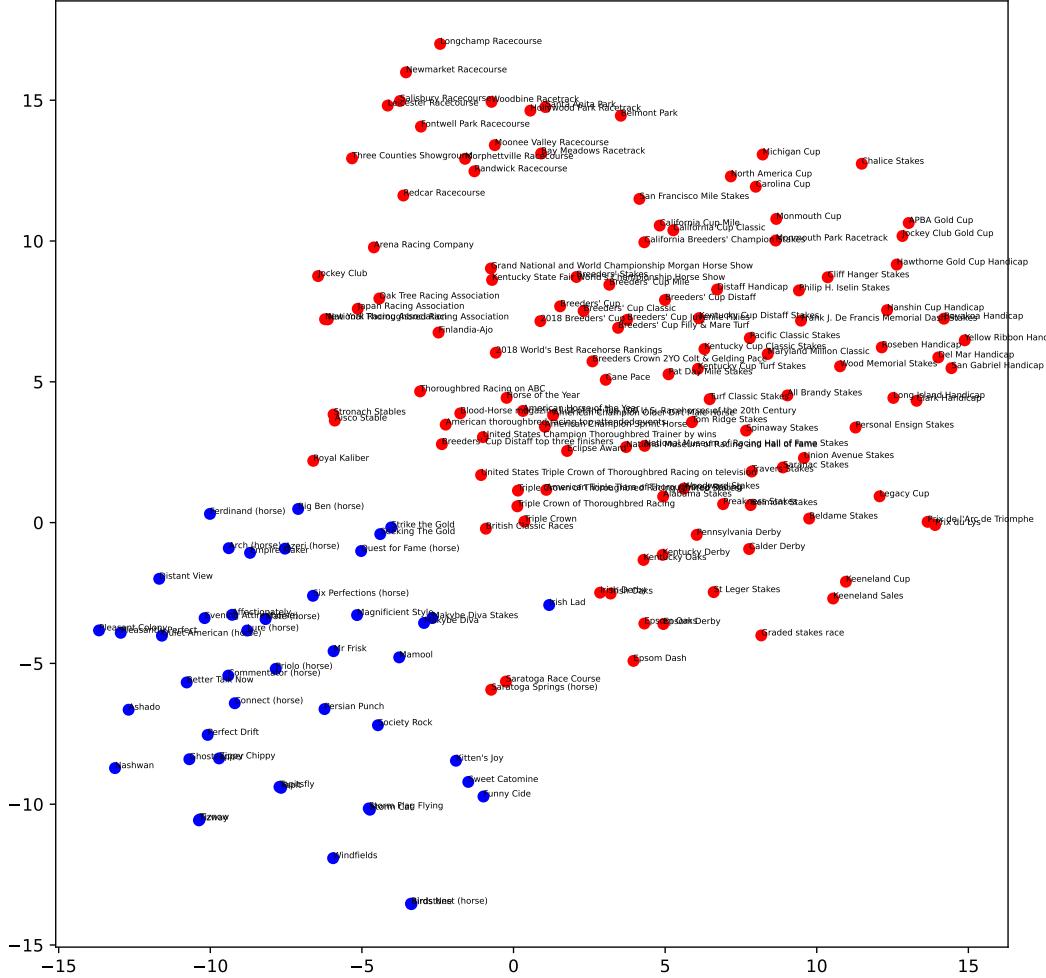


Figure 6: Famous racing horses merged with other horse racing entities.

dynamics, though with DIS exerting the stronger influence in most scenarios.

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.

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 5.

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 **SFSe**s. Overall, our parameter study reveals that achieving uniform semantic shift distribution depends primarily on the standard deviation multiplier rather than the minimum window size con-

Table 5: Performance metrics across standard multipliers.

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

figuration. The multiplier of 3.0 effectively prevents later iterations from dominating the analysis while maintaining sensitivity to meaningful semantic changes throughout the clustering process.

More Analysis on Hierarchical Structure

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 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 8. This follows a logical hierarchical structure, from more specific categories to broader ones. Yet, as shown in Figure 6 (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 the need for cautious interpretation: semantic hierarchies derived from embeddings may not faithfully mirror human conceptual structures and should be analyzed with an awareness of the model’s inductive biases.