computers

MDPI

Article

Set-Word Embeddings and Semantic Indices: A New Contextual Model for Empirical Language Analysis

Pedro Fernández de Córdoba [1], Carlos A. Reyes Pérez [1], Claudia Sánchez Arnau [2] and Enrique A. Sánchez Pérez [1,*]

[1] Instituto Universitario de Matemática Pura y Aplicada, Universitat Politècnica de València, 46022 València, Spain; pfernandez@mat.upv.es (P.F.d.C.); careyper@doctor.upv.es (C.A.R.P.)
[2] E.T.S. Ingeniería, Universitat de València, 46100 Valéncia, Spain; sanar4@alumni.uv.es
* Correspondence: easancpe@mat.upv.es; Tel.: +34-963877000

Abstract: We present a new word embedding technique in a (non-linear) metric space based on the shared membership of terms in a corpus of textual documents, where the metric is naturally defined by the Boolean algebra of all subsets of the corpus and a measure μ defined on it. Once the metric space is constructed, a new term (a noun, an adjective, a classification term) can be introduced into the model and analyzed by means of semantic projections, which in turn are defined as indexes using the measure μ and the word embedding tools. We formally define all necessary elements and prove the main results about the model, including a compatibility theorem for estimating the representability of semantically meaningful external terms in the model (which are written as real Lipschitz functions in the metric space), proving the relation between the semantic index and the metric of the space (Theorem 1). Our main result proves the universality of our word-set embedding, proving mathematically that every word embedding based on linear space can be written as a word-set embedding (Theorem 2). Since we adopt an empirical point of view for the semantic issues, we also provide the keys for the interpretation of the results using probabilistic arguments (to facilitate the subsequent integration of the model into Bayesian frameworks for the construction of inductive tools), as well as in fuzzy set-theoretic terms. We also show some illustrative examples, including a complete computational case using big-data-based computations. Thus, the main advantages of the proposed model are that the results on distances between terms are interpretable in semantic terms once the semantic index used is fixed and, although the calculations could be costly, it is possible to calculate the value of the distance between two terms without the need to calculate the whole distance matrix. "Wovon man nicht sprechen kann, darüber muss man schweigen". Tractatus Logico-Philosophicus. L. Wittgenstein.

Keywords: word embedding; semantic projection; set metric; Lipschitz function; semantic index

MSC: 68T35; 46B85; 28C15

Academic Editor: Ming Liu

Received: 3 December 2024
Revised: 10 January 2025
Accepted: 17 January 2025
Published: 20 January 2025

Citation: Fernández de Córdoba, P.; Reyes Pérez, C.A.; Sánchez Arnau, C.; Sánchez Pérez, E.A. Set-Word Embeddings and Semantic Indices: A New Contextual Model for Empirical Language Analysis. *Computers* **2025**, *14*, 30. https://doi.org/10.3390/computers14010030

1. Introduction

Nowadays, the fast improvement of natural language processing (NLP) can be seen in how all the applications of this discipline have reached impressive heights in the handling of natural language in many applied fields, ranging from the translation from one language to another, to the generalization of the use of generative artificial intelligence [1–4]. A fundamental tool for these applications is the so-called word embeddings. Broadly speaking,

word embeddings allow representations of semantic structures as subsets of linear spaces endowed with norms. The distance provided by such a norm has the meaning of measuring the relations among terms in such a way that two items being close means that they are close with respect to their meaning. The main criticism of this technique when it was first introduced is that, once a word embedding is created, its representation remains static and cannot be adapted to different contexts [1,3]. This is a problem for applications, as there are words whose meaning clearly depends on the context: take the word "right", for example. However, since the late 2010s, these models have evolved towards more flexible architectures, incorporating structural elements and methods that allow the integration of contextual information [5–7].

On the other hand, let us recall that the technical tools for natural language processing have their roots in some broad ideas about the semantic relationships between meaningful terms, but they are essentially computed from "experimental data", that is, searching information in concrete semantic contexts provided by different languages. Thus, one of the basic ideas underlying language analysis methods could be called the "empirical approach", which states that the meaning of a term is reflected in the words that appear in its contextual environment. An automatic way of applying this principle and transforming it into a mathematical rule is to somehow measure which sets of semantically non-trivial documents in a given corpus share two given terms. This is the essence of the method based on the so-called semantic projections used in this paper [8,9].

The aim of this article is to present a new mathematical formalism to support all NLP methods based on the general idea of contextual semantics. Thus, we propose an alternative to the standard word embeddings, which is based on a different mathematical structure (instead of normed spaces) developed using algebras of subsets endowed with the usual set-theoretic operations (Figure 1). In a given environment and for a given specific analytical purpose, the meaning of a term may be given by the extent to which it shares semantic context with other terms, and this can be measured by means of the relations among the sets representing all the semantic elements involved [10,11]. The value of the expression "the term v shares the semantic environment with the term w" can be measured by a real number in the interval $[0, 1]$ given by a direct calculation, which is called the semantic projection of v onto w. This representation by an index belonging to $[0, 1]$ is also practical when we think about the interpretation of what a semantic projection is, since it can be understood as a probability—thus facilitating the use of Bayesian tools for further uses of the models—as well as a fuzzy measure of how a given term belongs to the meaning of another term—thus facilitating the use of arguments related to fuzzy metrics.

Thus, we present here a mathematical paper, in which we define our main concepts and obtain the main results on the formal structure that supports empirical approaches to NLP based on the analysis of a corpus of semantically non-trivial texts, and how meaningful terms share their membership in these texts. The main outcome of the paper is a theoretical result (Theorem 2) which aims to demonstrate that our proposed method is universal, in the sense that any semantic structure which is representable through conventional word embeddings can also be represented as a set-word embedding, and vice versa. However, our approach to constructing representations of semantic concepts, while limited to the idea that semantic relationships are expressed through distances, is much more intuitive. This is because it is grounded in the collection of semantic data associated with a specific context. Thus, the key advantage of our model lies in its adaptability to contextual information, allowing it to be applied across various semantic environments.

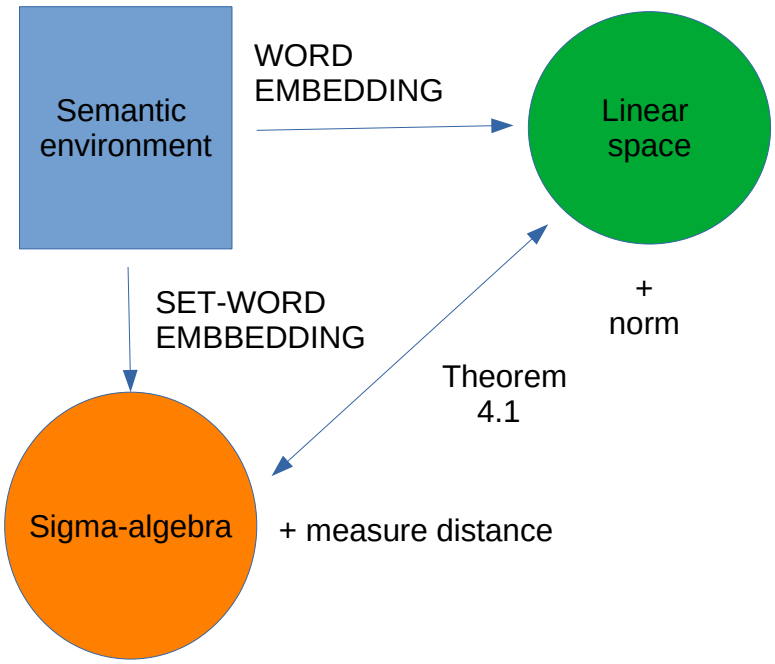

Figure 1. Word embedding versus set-word embedding.

This paper is structured as follows. In the subsequent subsections of this Introduction, we outline the specific context of our research, covering standard word embeddings, families of subsets in classical NLP models, and relevant mathematical tools. We will also give an overview of the word embedding methods that have appeared recently and are closely related to the purpose we bring here. After the introductory Sections 1 and 2, Section 3 introduces the fundamental mathematical components of our semantic representations, demonstrating key structural results and illustrating how these formal concepts relate to concrete semantic environments. To ensure a comprehensive understanding of our modeling tools, Section 4 presents three particular examples and applications. Following this motivational overview, Section 5 details the main result of the paper (Theorem 2), explaining its significance and providing further examples. Section 6 is devoted to the discussion of the result, built around a concrete example that allows the comparison of our tool with, firstly, the Word2Vec word embedding, and in a second part, with a more recent selection of word embeddings that we have considered suitable for their relation to our ideas. Finally, Section 7 offers our conclusions.

2. Related Work

In this section, we provide an overview of the relevant background literature and introduce the key concepts that will be utilized throughout the remainder of the paper.

2.1. Word Embeddings and Linear Spaces

Let us start this part of basic concepts by explaining what a word embedding is and showing in broad strokes the "state of the art" regarding this important NLP tool. Word embeddings have revolutionized natural language processing (NLP) by representing sets of words (in general, semantic items) as dense subsets of vectors in a high-dimensional linear space, capturing semantic and syntactic relationships. These representations operate on the principle that words appearing in similar contexts tend to have similar meanings. The learning process of word embeddings using neural networks involves initializing word vectors randomly and then training the network to predict words based on their context or vice versa. This approach has evolved significantly over time, with several key

models marking important milestones. Word2Vec [4] utilized shallow neural networks with a single hidden layer, proposing two efficient architectures: Continuous Bag-of-Words (CBOW) and Skip-gram. The CBOW architecture predicts a target word given its context, using the average of the context word vectors as input to a log-linear classifier. Conversely, the Skip-gram model predicts the context words given a target word, effectively inverting the CBOW architecture. Although the algorithms become rather technical, the main idea behind these models is to obtain a linear representation of the semantic space in which distances (measured by means of norms) represent meaning similarity. Therefore, these models focused on capturing word relationships in vector space and introduced innovations such as negative sampling and hierarchical softmax for efficient training. Building upon this foundation, GloVe was developed [5], which combined global matrix factorization with local context window methods, explicitly modeling the ratio of co-occurrence probabilities. GloVe's architecture involves a weighted least squares regression model that learns word vectors by minimizing the difference between the dot product (scalar product of the underlying Euclidean space) of word vectors and the logarithm of their co-occurrence probability.

A paradigm shift occurred with the introduction of BERT [2] and GPT [6], which moved from static word embeddings to contextual embeddings and introduced the "pre-training and fine-tuning" paradigm. BERT utilized a bidirectional transformer architecture for capturing rich contextual information. The transformer architecture, first proposed in [7], relies on self-attention mechanisms to process input sequences in parallel, allowing for more efficient training on large datasets. BERT's pretraining involves two tasks: Masked Language Modeling (MLM), where the model predicts masked tokens in a sentence, and Next Sentence Prediction (NSP), where it predicts if two sentences are consecutive in the original text. This bidirectional approach allows BERT to capture context from both left and right of a given word. On the other hand, GPT demonstrated the power of unidirectional language modeling at scale, using only the decoder portion of the transformer architecture. GPT's architecture processes text from left to right, predicting each token based on the previous tokens, which allows for efficient generation of coherent text.

These advancements reflect key trends in the field, including increasing model size and complexity, a shift from local to global context, evolution from static to dynamic embeddings, and a growing emphasis on unsupervised pretraining on large corpora. The progression from word-level to subword-level tokenization methods, such as WordPiece used in BERT and byte-pair encoding used in GPT, has further enhanced the capability of these models to handle diverse vocabularies and capture nuanced semantic information.

In recent times, more advanced methods have been introduced that affect the way data are treated to achieve better information processing to build word embeddings, as well as to improve and diversify applications. NLP models are a highly active research area, with significant advancements made each year. Methods such as LDA, LSA, PLSA, Bag of Words, TF-IDF, Word2Vec, GloVe, and BERT, among other natural language processing applications (see, for example, ref. [12] for an overview of these models), have become foundational in the field, despite some of them being just eight years old. These paradigms continue to evolve, and the range of applications has expanded so much that it has become increasingly difficult to track which models are leading in terms of performance. However, recent advances show a clear shift towards more contextualized and formal structures in some new techniques, with a greater focus on how information is represented and processed [13]. Despite the diversity in approach, most of these models rely on high-dimensional Euclidean spaces as their core structure. Additionally, methods for introducing complementary information into models are rapidly evolving, with novel approaches emerging for improved learning [14,15]. A comprehensive analysis of these

new methodologies is beyond the scope of this paper, but understanding the direction in which these techniques are evolving is essential for our work.

But these modifications have not affected the main representational structure underlying the word embeddings, i.e., the mathematical environment chosen to establish a conceptual isomorphism between a semantic context and a formal structure: the linear normed space. As will be seen, in this paper we propose an alternative mathematical context, provided by Boolean algebras of subsets endowed with metrics specially designed for the representation of semantic structures.

2.2. Classes of Subsets as Natural Language Models

As outlined in the previous section, the methods for representing semantic relations through word embeddings fundamentally employ a linear space representation. This is achieved by mapping a set of tokens with significant semantic value into a normed space. Initially, these items are embedded as independent vectors within a high-dimensional linear space. Various procedures, such as encoder–decoder methods (see [3]), then enable the reduction of the representing space's dimensions while also adjusting the (Euclidean) distances to reflect the semantic proximity among the terms. This approach creates a highly effective framework for applications. However, conceptually, representations based on classes of sets—capturing the notion that semantic relations among terms can be expressed as inclusions, unions, and intersections—are arguably more appealing. This perspective aligns more closely with our intuitive understanding of semantic relations.

It is worth noting that this idea is not new in the field of natural language processing. For instance, Zadeh's original work, which laid the groundwork for fuzzy set theory, fundamentally addressed these types of mathematical structures [10]. The core premise of fuzzy set theory is that membership in a set is not merely a Boolean variable but, rather, a continuous variable. In other words, a semantic term (such as a word, token, or item) belongs to a given concept to a degree represented by an index with values ranging from 0 to 1, indicating the intensity of that relationship. For example, while "lion" belongs to the concept of "animal", "healthy" relates to "happiness" at a certain (high) coefficient, which does not reach 1, since healthy individuals are not necessarily happy. Moreover, standard operations in vector spaces—such as addition and scalar multiplication—do not accurately reflect the properties of semantic relationships between words. While some attempts to bridge this gap have been rigorously explored in the literature (see [1,16,17]), they have yet to fully align with intuitive semantic notions.

This context presents a broader view: linear space word embeddings are highly effective operationally, yet they often lack the intuitive appeal found in using algebras of subsets as a mathematical foundation for NLP. The primary focus of this paper is to demonstrate that both representations are, in fact, equivalent. Specifically, every vector-valued word embedding can be interpreted as a set-word embedding, and vice versa. This is presented and proved in Theorem 2.

2.3. Mathematical Tools

Let us explain here some notions that will be necessary for the explanation of the ideas in this paper, that, as we said, is mainly of mathematical nature. We will use standard definitions and notations of set theory, measure theory, metric spaces, and Lipschitz functions.

Let us recall first what a $\sigma-$algebra of subset is. Consider a set X. A $\sigma-$algebra is a class $\Sigma \subset \mathcal{P}(X)$ of subsets of X that is closed under countable unions and complements of elements of Σ. That is, if $(A_n)_n \subseteq \Sigma$, then $\cup_n A_n \in \Sigma$, and $X \setminus A = A^c$ belongs to Σ too if $A \in \Sigma$. As a consequence, finite intersections of elements also belong to the $\sigma-$algebra.

A (countably additive real) positive measure on a $\sigma-$algebra $\Sigma \to \mathbb{R}^+$ is a function that is countably additive when acting on countable families if pairwise disjoint sets, and allows us to define integrals of real valued measurable functions. The examples that we show in the paper are constructed using what are referred to as (strictly positive) atomic measures μ, which means that $\mu(\{v\}) > 0$ for every set with a single point of X.

Let us provide now the basics on the metric spaces we use as mathematical support. Consider a set X and a positive real function $d : X \times X \to \mathbb{R}$. If it satisfies that for every $x, y, z \in X$, we have that (1) $d(x,y) = 0$ if and only if $x = y$, (2) $d(x,y) = d(y,x)$, and (3) $d(x,z) \leq d(x,y) + d(y,z)$, it is said that d is a metric, and (X,d) is a metric space. In this paper we will consider a special type of metric space for which the elements are subsets of a given set belonging to a $\sigma-$algebra Σ. This will be endowed with particular metrics to become a metric space, which will be fundamental for our representation of the semantic relations.

Given a metric space (X,d) with a distinguished point 0, the Arens–Eells space $AE(X)$ (also called the free space) is defined as the vector space of all the molecules. We will define it later; let us first introduce some elementary notions. Consider another metric space (Y,ρ). A function $f : X \to Y$ is called Lipschitz if there exists a constant $K > 0$ such that

$$\rho\big(f(x), f(y)\big) \leq K\, d(x,y), \quad x, y \in X.$$

The relation between metric spaces and normed linear spaces is fundamental for this paper, so let us center the attention in which (Y,ρ) is a Banach space $(E, \|\cdot\|)$. The class of Lipschitz mappings from X to E that vanish at 0 is a Banach space, denoted by $Lip_0(X,E)$. The Lipschitz norm $Lip(T)$ for a map $T \in Lip_0(X,E)$, is the smallest constant $C \geq 0$ such that

$$\|T(x) - T(y)\| \leq C d(x,y), \quad x, y \in X.$$

The case in which E is the Euclidean real line $\mathbb{R}$, the notation $X^{\#} = Lip_0(X) = Lip_0(X,\mathbb{R})$, is often used. This is space is known as the Lipschitz dual of X. Let us introduce now the so-called Arens–Eells space of a metric space (X,d), denoted by $AE(X)$ [18], which is a fundamental piece of the construction presented in this document. Sometimes $AE(X)$ is referred to as the free space associated with X, and has the main feature that it is the predual of $Lip_0(X)$, that is,

$$(AE(X))^* = Lip_0(X).$$

The essential vectors of $AE(X)$ (from which the space is generated) are the so-called molecules on X. A molecule is a real-valued function m on X with finite support, satisfying $\sum_{x \in X} m(x) = 0$. This means that all of them can be written as finite sums and differences of simple molecules that are described below. If $x, x' \in X$, the molecule $m_{xx'}$ is given by $m_{xx'} = \chi_{\{x\}} - \chi_{\{x'\}}$, where χ_S denotes the characteristic function of the set $S = \{x, x'\}$, that is, the functions that equals 1 at S and 0 out of S. The space of all the molecules is the real linear space $\mathcal{M}(X)$.

A norm can be given for this space. If $m \in \mathcal{M}(X)$ we consider a representation as $m = \sum_{j=1}^{n} \lambda_j m_{x_j x'_j}$, and its norm is defined as

$$\|m\|_{\mathcal{M}(X)} = \inf\left\{ \sum_{j=1}^{n} |\lambda_j| d(x_j, x'_j),\ m = \sum_{j=1}^{n} \lambda_j m_{x_j x'_j} \right\},$$

where infimum is computed over all possible representations of m as the one written above. The norm completion of $(\mathcal{M}(X), \|\cdot\|_{\mathcal{M}(X)})$ is called the Arens–Eells space, $AE(X)$. The function $\iota : X \to AE(X)$ given by $\iota(x) = m_{x0}$ is an isometric embedding of X into $AE(X)$.

3. Set-Word Embeddings: The Core

Word embeddings are understood to take values in vector spaces. However, as we explain in the Introduction, our contextual approach to automatic meaning analysis suggests that perhaps a different embedding procedure based on set algebra would better fit the idea of what a word embedding should be. Let us show how to achieve this.

We primarily consider a set of words W, and a $\sigma-$algebra $\Sigma(S)$ of subsets of another set S.

Definition 1. *A set-word embedding is a one-to-one function* $\iota : W \hookrightarrow \Sigma(S)$,

In this setting, it is natural to consider the context provided by the Jaccard (or Tanimoto) index, which is a similarity measure that has applications in many areas of machine learning that is given by the expression

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}, \tag{1}$$

where A and B are finite subsets of a set W. It is also relevant that the complement of this index,

$$D(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}, \tag{2}$$

is a metric [19] called the Jaccard distance. The general version of this measure is the so-called Steinhaus distance, which is given by

$$S_\mu(A,B) = \frac{\mu(A \cup B) - \mu(A \cap B)}{\mu(A \cup B)}, \tag{3}$$

where μ is a positive finite measure on a $\sigma-$algebra Σ, and A, B are measurable sets ([20], S.1.5). The distance proposed in [21] is a generalization of these metrics.

Let us define a semantic index characteristic of the embedding i, which is based on a kind of asymmetric version of Jaccard's index. In our context, and because of the role it plays for the purposes of this article, we call it a semantic index of one meaningful element onto another in a given context. In the rest of this section, we assume that μ is a finite positive measure on a set S, acting on a certain $\sigma-$algebra $\Sigma(S)$ of subsets of S.

Definition 2. *Let* $A, B \subset \Sigma(S)$. *The semantic index of* B *on* A *is defined as*

$$P_A(B) := \frac{\mu(A \cap B)}{\mu(A)}. \tag{4}$$

Therefore, once a measurable set A *is fixed, the semantic index is a function* $P_A : \Sigma \to \mathbb{R}^+$.

Those functions are called semantic projections in [9]. Roughly speaking, this rate provides information about which is the "proportion of the meaning" of A that is explained/shared by the meaning of B. But this is a mathematical definition, and the word "meaning" associated with a subset $A \in \Sigma$ is just the evaluation of the measure μ on it. As usual, μ plays the role of measuring the size of the set A with respect to a fixed criterion.

In [9], the definition of semantic projection was made in order to represent how a given conceptual item (canonically, a noun) is represented in a given universe of concepts. With the notation of this paper, given a term A and a finite universe of words $\mathcal{U}$, the semantic projection $(\alpha_u(A))_{u \in \mathcal{U}}$ is defined as a vector ([9], S.3.1) in which each coordinate is given by

$$\alpha_u(A) = p(A)|_B = \frac{|A \cap B|}{|A|}, \tag{5}$$

where B is the subset that represents the noun u of the universe in the word embedding ([9], S.3.1). These coefficients essentially represent a particular case of the same idea as the semantic index defined above, as for μ being the counting measure,

$$\alpha_u(A) = p(A)|_B = \frac{|A \cap B|}{|A|} = P_A(B). \tag{6}$$

We call it semantic index instead of semantic projection to avoid confusion.

Let us now define the framework of non-symmetric distances for the model.

Definition 3. *Let $\Sigma(S)$ be a $\sigma-$algebra of subsets of a given set S. Let μ be a finite positive measure on S. We define the functions*

$$q^L_\mu(A, C) := \mu(A \cap C^c), \quad q^R_\mu(A, C) := \mu(A^c \cap C), \quad A, B \in \Sigma(S), \tag{7}$$

as well as

$$q_\mu := q^L_\mu + q^R_\mu. \tag{8}$$

Let us see that the so-defined expressions provide the (non-symmetric) metric functions that are needed for our construction. As in the case of the Jaccard index mentioned above, there is a fundamental relationship between q_μ and P_B, which is given in (iii) of the following result.

Lemma 1. *Fix a σ-algebra $\Sigma(S)$ on S, and let μ be a finite positive measure on the measurable sets of S. Then,*

(i) *q^L_μ and q^R_μ are conjugate quasi-metrics on the set of the classes of $\mu-$a.e. equal subsets in $\Sigma(S)$.*
(ii) *q_μ is a metric on this class.*
(iii) *For every $A, B \in \Sigma(S)$,*

$$q^L_\mu(A, B) = \mu(A)\Big(1 - P_A(B)\Big) \quad \textit{and} \quad q^R_\mu(A, B) = \mu(B)\Big(1 - P_B(A)\Big). \tag{9}$$

Proof. Note that, for every $A, B, C \in \Sigma(S)$,

$$A \cap C^c = (A \cap C^c \cap B^c) \cup (A \cap C^c \cap B) \subset (A \cap B^c) \cup (B \cap C^c).$$

Consider

$$q^L_\mu(A, C) = \mu(A \cap C^c) \quad \text{and} \quad q^R_\mu(A, C) = \mu(A^c \cap C).$$

Clearly,

$$q^L_\mu(A, B) = q^R_\mu(B, A).$$

and so both expressions define conjugate functions. Then,

$$q^L_\mu(A, C) := \mu(A \cap C^c) \leq \mu(A \cap B^c) + \mu(B \cap C^c) = q^L_\mu(A, B) + q^L_\mu(B, C).$$

and the same happens for q^R_μ, and so both q^L_μ and q^R_μ are quasi-pseudo-metrics. To see that they are quasi-metrics, we have to prove (ii), that is, that the symmetrization

$$q = q^L_\mu + q^R_\mu$$

is a distance (on the class of $\mu-$a.e. equal sets). But if $q_\mu(A, B) = 0$, we have that both

$$\mu(A \cap B^c) = 0 \quad \text{and} \quad \mu(A^c \cap B) = 0.$$

This can only happen if $\mu(A) = \mu(A \cup B) = \mu(B)$, and so $A = B$ $\mu-$a.e. As q_μ is symmetric and satisfies the triangle inequality, we obtain (ii), and so (i).

(iii) For every $A, B \in \Sigma(S)$,

$$q_\mu^L(A, B) = \mu(A \cap B^c) = \mu(A) - \mu(A \cap B) = \mu(A)\Big(1 - P_A(B)\Big),$$

and $q_\mu^R(A, B) = q_\mu^L(B, A) = \mu(B)\Big(1 - P_B(A)\Big)$.

□

It can be seen that the metric q_μ that we have defined coincides with the so-called symmetric difference metric (the Fréchet–Nikodym–Aronszyan distance, which is given by $d(A, B) = \mu\big((A \cup B) \setminus (A \cap B)\big)$ (see [20], S.1.5). However, our understanding of relations between sets is essentially non-symmetric, since semantic relations are, in general, not symmetric: to use a classical example, in a set-based word representation, it makes sense for the word "king" to contain the word "man", but not vice versa. It seems natural that this is reflected in the distance-related functions for comparing the two words. This is why we introduce metric notions using quasi-metrics as primary distance functions. It must be said that, in the context of the abstract theory of quasi-metrics, the distance $q_{\mu,\max} = \max\{q_\mu^L, q_\mu^R\}$ is used instead of q_μ (see, for example, [22]). It is equivalent to q_μ and plays exactly the same role as $q_{\mu,\max}$ from the theoretical point of view: we use q_μ for its computational convenience, as we will see later.

Summing up the information presented in this section, we have that the fundamental elements of a set-word embedding are the following.

- A set of terms (words, short expressions, ...) W, on which we want to construct our contextual model for the semantic relations among them.
- A finite measure space $(S, \Sigma(S), \mu)$, in which the $\sigma-$algebra of subsets have to contain the subsets representing each element of W, and S can be equal to W or not.
- The word embedding itself: an injective map $\iota : W \hookrightarrow \Sigma(S)$, that is well defined as a consequence of the requirement above.
- The quasi-metrics q_μ^L and q_μ^R, which, together with the metric q_μ, give the metric space $(\Sigma(S), q)$ that supports the model.

Remark 1. *In principle, it is required that μ be positive. However, this requirement is not necessary in general for the construction to make sense, and it is even mandatory to extend the definition for general (non-positive) finite measures, as will be shown in the second part of the paper. What is actually required is that the symmetric differences of the elements of the range of the set-word embedding have positive measure. We will see that this is, in fact, weaker than being a positive measure.*

The next step is to introduce a systematic method for representing the features we need to use about the elements of W that would help to enrich the mathematical structure of the representations. This is achieved in the next subsection.

3.1. Features as Lipschitz Functions on Set-Word Embeddings

By now we have already introduced the basic mathematical tools to define a set-word embedding. The method we used to perform it made them derived from a standard index, which we call the semantic index of a word $w_2 \in W$ with respect to another word $w_1 \in W$. It has a central meaning in the model. We prove that it is a Lipschitz function in Theorem 1 below. In a sense, we think of it as a canonical index that shows the path to model any new features we want to introduce into the representation environment.

Our idea is also influenced by the way some authors propose to represent the "word properties" for the case of vector-valued word embeddings $i : W \hookrightarrow \mathbb{R}^n$ (we use "i" for this map to follow the standard notation for inclusion maps, but the reader should be careful because it can sometimes be confusing). Linear operations on support vector spaces are widely used for this purpose [1,8,16,23,24]. Thus, linear functions are sometimes used to define what are called semantic projections, which can be written as translations of linear functionals of the dual space V^* of the vector space V. For example, to represent the feature "size" in a word embedding of animals, it is proposed in [8] to use a regression line, and this defines what they call a "semantic projection". The same idea is used in [9]; in this case, the semantic projections are given by the coordinate functionals of the corresponding vector-valued representation.

This opens the door to our next definition. But in our case, we do not have any kind of linear structure, so we have to understand our functions as elements of the "dual" of the metric space $(\Sigma(S), q)$. If we take the empty set $\emptyset \in \Sigma(S)$ as a distinguished point 0 in $(\Sigma(S), q)$, this dual space is normally defined as the Banach space of real-valued Lipschitz functions that are equal to 0 at $\emptyset$, $\big(Lip_0(\Sigma(S)), Lip(\cdot)\big)$. The norm is defined as $Lip(\varphi) + |\varphi(\emptyset)|$ if we do not assume that φ is 0 at $\emptyset$, and then $Lip(\Sigma(S))$ is also a Banach space (see, e.g., [25,26] for more information on these spaces).

Definition 4. *An index representing a word-feature on the set-word embedding is a real-valued Lipschitz function $\varphi : (\Sigma(S), q) \to \mathbb{R}$. In the case that there is a constant $q_\mu > 0$ such that φ satisfies also*

$$q_\mu(A, B) \leq Q \ \big(\varphi(A) + \varphi(B)\big) \quad \textit{for all} \quad A, B \in \Sigma(S), \tag{10}$$

we will say that φ is a $q_\mu-$Katětov (or q_μ-normalized) function (index). We write $Kat(\varphi)$ for the infimum of such constants Q.

Lipschitz functions with $Lip(\varphi) = 1$ and satisfying the second condition for $q_\mu = 1$ are often called Katětov functions (see, for example, ([20], S.1)). Under the term q_μ-normalized index, the second requirement in this definition is used in [27] in the context of a general index theory based on Lipschitz functions. More information on real functions satisfying both requirements above can be found in this paper.

A standard case of $q_\mu-$Katětov Lipschitz functions (called sometimes Kuratowski functions, standard indices in [27]) are the ones given by $\Sigma(S) \ni B \mapsto \varphi_A(B) := q_\mu(A, B)$ for a fixed $A \in \Sigma(S)$. Indeed, note that for $A \in \Sigma(S)$,

$$|\varphi_A(B) - \varphi_A(C)| = |q_\mu(A, B) - q_\mu(A, C)|$$

$$\leq q_\mu(B, C) \leq q_\mu(A, B) + q_\mu(A, C) = \varphi_A(B) + \varphi_A(C)$$

for every $B, C \in \Sigma(S)$. This, together with a direct computation using $\varphi_A(A) = 0$, show that $Lip(\varphi_A) = 1$ and $Kat(\varphi_A) = 1$. Indeed, note that under the assumption that S is finite, we can easily see the following:

(1) For every Lipschitz function $\varphi : \Sigma(S) \to \mathbb{R}$, we can find a real number r, a non-negative Lipschitz function $\varphi^* : \Sigma(S) \to \mathbb{R}^+$, and a set $A \in \Sigma(S)$ such that $\varphi^* = \varphi + r$ and $\varphi(A) = 0$.
(2) $Lip(\varphi) = Lip(\varphi^*)$ and $Lip(\varphi^*) \cdot Kat(\varphi^*) \geq 1$, that is, φ can be translated to obtain a non-negative function obtaining the value 0 at a certain set and preserving the Lipschitz constant; also, a direct calculation, as per the one above for φ_A, shows that the product of $Lip(\varphi^*)$ can never be smaller than 1.

The functions described above are a particular application of the so-called Kuratowski embedding [28], which is defined as the map $k : X \hookrightarrow C_b(X)$ given by the formula $k(x)(y) = d(x,y) - d(x_0,y)$, $x, y \in X$, where (X,d) is a metric space, $C_b(X)$ is the Banach space of all bounded continuous real-valued functions on X, and x_0 is a fixed point of X. If we take $x = x_0$, we obtain our functions.

The above notion motivates the following definition of compatibility index. For a Lipschitz index φ, the compatibility index given below gives a measure of how close φ is to the given metric in space, such that a small value (it always has to be greater than 1) indicates that there is a close correlation between the relative values of φ in two generic subsets A and B and the value of the distance $q_m u(A,B)$. Conceptually, a small value of $Com(\varphi)$ represents the (desired) deep relationship between the semantic index and the metric in the space, thus functioning as a quality parameter for the model.

Definition 5. *Let $\varphi : \Sigma(S) \to \mathbb{R}$ be a Lipschitz $q_\mu-$Katětov positive function attaining the value 0 only at one set $A \in \Sigma(S)$. We call the constant $Com(\varphi)$ given by*

$$Com(\varphi) = Lip(\varphi) \cdot Kat(\varphi) \tag{11}$$

the compatibility index of φ with respect to the metric space $(\Sigma(S), q_\mu)$. We have already shown that $Com(\varphi) \geq 1$.

The importance of the compatibility index in the proposed metric model is clear: the smaller the constant $Com(\varphi)$, the better the index φ fits the metric structure. That is, whenever we use φ to model any semantic feature in the context of the set-word embedding $\iota : W \hookrightarrow \Sigma(S)$, we can expect the feature represented by φ to "follow the relational pattern between the elements of W" as closely as $Com(\varphi)$ is small. Let us illustrate these ideas with the next simple example.

Example 1. *Two different features concerning a set of nouns (for example, two adjectives) do not necessarily behave in the same fashion. For example, take $W = \{w_1 = lion, w_2 = horse, w_3 = elephant\}$, and two properties: φ_1, which represents the adjective "big", and φ_2, which represents "fierceness". We can define the degree of each property by*

$$\varphi_1(w_1) = 1, \quad \varphi_1(w_2) = 2 \quad \varphi_1(w_3) = 3,$$

and

$$\varphi_2(w_1) = 3, \quad \varphi_2(w_2) = 0 \quad \varphi_2(w_3) = 1.$$

Consider the trivial set-word embedding $\iota : W \hookrightarrow D_3$, where $D_3 = \{1,2,3\}$, $\iota(w_i) = i$ for $i = 1,2,3$, and μ is the counting measure on the $\sigma-$algebra of subsets of D. Then,

$$q_\mu(i,j) = \mu(\{i,j\} \setminus \emptyset) = 2, \quad i,j = 1,2,3 \quad \text{if } i \neq j, \quad \text{and} \quad 0 \text{ otherwise.}$$

Obviously, both indices are Lipschitz, and $Lip(\varphi_1) = 1$ and $Lip(\varphi_2) = 3/2$. On the other hand, both of them are Katětov functions with $Kat(\varphi_1) = 2/3$ and $Kat(\varphi_2) = 2$.

Let us compute the compatibility indices associated with both φ_1 and φ_2. First, note that φ_1 has to be translated to attain the value 0; let us define $\varphi_1^* = \varphi_1 - 1$. We have that $Kat(\varphi_1^*) = 2$, so

$$Com(\varphi_1^*) = Lip(\varphi_1^*) \cdot Kat(\varphi_1^*) = 1 \cdot 2 = 2,$$

and

$$Com(\varphi_2) = Lip(\varphi_2) \cdot Kat(\varphi_2) = (3/2) \cdot 2 = 3.$$

Following the explanation we have given for Com, we have that φ_1 better fits q_μ than φ_2.

Note that we cannot expect any kind of linear dependence among the representation provided by ι and the functions representing these properties. For example, the index that concerns the size φ_1 can be given by the line $\iota(w_i) = i$, while the second one, φ_2, does not satisfy any linear formula. In fact, it does not make sense to try to write it as a linear relation, as there is no linear structure on the representation by the metric space $(\Sigma(D_3), q_\mu)$.

From the conceptual point of view, this is the main difference of our purpose of set-word embedding versus the usual vector-word embedding. The union or intersection of two terms has a semantic interpretation in the model: the semantic object created by considering two terms together in the first case, and the semantic content shared by the two terms, respectively. However, the addition of two terms in their vector space representations or the multiplication of the vector representation of a given word by a scalar have dubious meanings in the models, although they are widely used [8,17,23].

Let us show now that the semantic index P_A is a Lipschitz index, and even $1-$Lipschitz, and $q_\mu-$Katětov with $Kat(P_A) = 1$ for μ being a probability measure and $A = S$. The main idea underlying the following result, which is fundamental to our mathematical construction, is the existence of a deep relationship between semantic indexes and the metric defined by a given measure μ. Essentially, this means that the model can be structured around the notion of semantic index, which has the clear role of a measure of the shared meaning between two terms, and an associated metric, which allows the comparison and measurement of distances between semantic terms.

Theorem 1. *For every $A \in \Sigma(S)$, the function $\Sigma \ni D \mapsto P_A(D) \in \mathbb{R}$ is Lipschitz with constant $Lip(P_A) \leq 1/\mu(A)$. That is,*

$$\Big|P_A(C) - P_A(B)\Big| \leq \frac{1}{\mu(A)}\, q_\mu(C, B), \qquad C, B \in \Sigma(S). \tag{12}$$

Moreover, for $A = S$, we also have

$$q_\mu(C, B) \leq \mu(S)\,\Big(P_S(C) + P_S(B)\Big), \quad C, B \in \Sigma(S), \tag{13}$$

and then $Com(P_S) = 1$.

Proof. As $\mu(A)\big(1 - P_A(B)\big) = q^L_\mu(A, B)$, taking into account that

$$q^L_\mu(A, B) - q^L_\mu(A, C) \leq q^L_\mu(C, B),$$

(the same for q^R_μ), we have that

$$\begin{aligned} P_A(C) - P_A(B) &= (1 - P_A(B)) - (1 - P_A(C)) \\ &= \frac{1}{\mu(A)}\Big(\mu(A)(1 - P_A(B)) - \mu(A)(1 - P_A(C))\Big) \\ &= \frac{1}{\mu(A)}\Big(q^L_\mu(A, B) - q^L_\mu(A, C)\Big) \leq \frac{1}{\mu(A)}\, q^L_\mu(C, B). \end{aligned}$$

The symmetric calculations give

$$P_A(B) - P_A(C) \leq \frac{1}{\mu(A)}\, q^L_\mu(B, C) = \frac{1}{\mu(A)}\, q^R_\mu(C, B).$$

Therefore,

$$\left|P_A(C) - P_A(B)\right| = \left(P_A(C) - P_A(B), P_A(B) - P_A(C)\right)$$
$$\leq \frac{1}{\mu(A)}\left(q_\mu^L(C,B) + q_\mu^R(C,B)\right) = \frac{1}{\mu(A)}\, q_\mu(C,B).$$

For the last statement, just note that

$$q_\mu(C,B) = \mu(C \cap B^c) + \mu(C^c \cap B)$$
$$\leq \mu(S)\left(\frac{\mu(C\cap S)}{\mu(S)} + \frac{\mu(B\cap S)}{\mu(S)}\right) = \mu(S)\left(P_S(C) + P_S(B)\right).$$

In particular, if μ is a probability measure, the function P_S above satisfies the inequalities

$$|P_S(C) - P_S(B)| \leq q_\mu(C,B) \leq P_S(C) + P_S(B), \quad C, B \in \Sigma(S),$$

and so it is a Katětov function such that $Com(P_S) = Lip(P_S) \cdot Kat(P_S) = 1 \cdot 1 = 1$. □

This result is fundamental to the interpretation of the model. Roughly speaking, it states that the main index on which we have relied conforms completely to the metric structure. Since μ is a probability measure, the semantic index P_A for the case $A = S$ represents the rate (the score per one) of information contained in any information set $B \in \Sigma(S)$, which is measured by $\mu(B)$. Thus, Theorem 1 ($Com(P_S) = 1$) means that this fundamental quantity absolutely fits the space $(\Sigma(S), q_m u)$, the main tool of our embedding procedure.

3.2. How to Apply a Set-Word Embedding for Semantic Analysis: An Example

To finish this section, let us sketch how the proposed model can be used for semantic analysis. We only intend to show some general ideas in this paper, and compare them with the ones that are usual tools in the context of the vector-word embeddings.

Let us focus our attention on a binary property that regards a set of words representing nouns of a certain language. Write $(\Sigma(S), q_\mu)$ for the metric space defined by all the subsets of S, with $(S, \Sigma(S), \mu)$ a probability measure space, and let $\iota : W \hookrightarrow \Sigma(S)$ the set-word embedding in which we base our model (note that this "ι" is not the usual "i" used before to denote a standard word embedding.) The set S could be, for example, a class of properties of the animals: average size, taxonomic distance, common color, eat grass or not…, and ι embeds every animal in the set of properties that it has. The measure μ quantifies the relevance in the model that each of the properties in S has.

The studied feature is described by the values 0 or 1; so we call the Lipschitz functional with values in $\{0,1\}$ representing the feature a classifier. For example, in a given set of animals W, the property of having two legs is represented by 0, and having four legs, by 1. Let us write ϕ for the Lipschitz map representing the property "having two/four legs". Let us consider a specific situation, and how to solve it using the proposed set-word embedding.

- Suppose that we know the value of the classifier ϕ at a subset $\iota(W_0) \subset \iota(W)$, but it is unknown out of W_0.
- The value of ϕ in $\iota(W) \setminus \iota(W_0)$ can be estimated using the evaluation on some terms of the original universe and then extending using a McShane–Whitney type formula. This extension provides the equation

$$\widehat{\phi}(a) = \frac{1}{2}\sup_{b\in S}\Big(\phi(b) - Lip(\phi)\, d(a,b)\Big) + \frac{1}{2}\inf_{b\in S}\Big(\phi(b) + Lip(\phi)\, d(a,b)\Big) \tag{14}$$

  which gives an estimate of the value of $\phi(a)$ with values in the interval $[0,1]$ for elements that do not belong to $\iota(W_0)$.

- Therefore, the structure of the metric space together with the explained Lipschitz regression tool provide information about the expected values of ϕ in the set $\iota(W)$. Since it takes values in the interval $[0,1]$, but not necessarily in $[0,1]$, we can interpret the values provided by $\widehat{\phi}$ in probabilistic terms: it gives the probability that a given element of $W \setminus W_0$ has two or four legs. Also, we can interpret it as the fuzzy coefficient of that element belonging to the set of four-legged animals.

4. Three Related Constructions

We show in this section three general frameworks in which the formalism explained can be used for semantic analysis. They are built from scratch, and we try to demonstrate with them how adaptable our method is by introducing set-theoretic arguments into the reasoning.

4.1. A Database of Dictionaries

We will show a specific construction for the semantic contextual analysis of a set of nouns N using a collection of dictionaries. Finding a good set of nouns for the analysis of a given semantic context would be an important preliminary work for a useful application of the abstract procedure described below. This problem is, in a sense, similar to the determination of a suitable set of words describing a given semantic environment using keywords [29]. Many techniques from natural language processing have been proposed to help in this process (Word Sense Disambiguation algorithms to obtain the best SynsetID; see [29] and the references therein). Something similar should be performed as a first step in our case. Wikipedia could be the source of information instead of our "collection of dictionaries".

Let us describe the constructive process step by step. Consider a fixed set of nouns, N, on a certain topic (for example, for Sociology of Migration: "culture", "acculturation", "assimilation", "autochthonous", ...,"process", "change","system", ...).

- Consider a family $\mathcal{D}$ of R dictionaries. Let us fix a word n in N, and consider the text appearing in the dictionary entry $k \in \mathcal{D}$ associated with n. Let us define the set of words v_n^k appearing in each of these texts.
- Now, take the sets $t_n := \{n\} \cup \left(\cup_{k=1}^{R} v_n^k \cap N \right)$, and define the class of subsets

$$T = \{t_n : n \in N\}.$$

  This is the basic set of elements we consider in this example. Each of them represents the set of nouns that are used, in all dictionaries, to give a definition of a given noun n, including the noun itself.
- Following our construction, take the $\sigma-$algebra $\Sigma(N)$ generated by the subsets of N, and note that $T \subset \Sigma(N)$. Define the word embedding

$$\iota : N \hookrightarrow \Sigma(N), \quad \iota(n) := t_n \in T.$$

  Thus, each noun is represented in the set-word embedding as the set of nouns appearing in all the entries of this word in the dictionaries of $\mathcal{D}$.

If the counting measure is used as underlying measure for the $\sigma-$algebra of all subsets of N, the quasi-distance q_μ^L between two nouns $n_1, n_2 \in N$ is given by

$$q_\mu^L(\iota(n_1), \iota(n_2)) = |\iota(n_1) \cap \iota(n_2)^c|, \tag{15}$$

that is, the number of all nouns appearing in the descriptions of n_1 in all the dictionaries of $\mathcal{D}$ that are *not* in the descriptions of n_2. Using also the dual definition of q_μ^R, we obtain the metric q_μ, which is their max-symmetrization.

The semantic index $P_{\iota(n_1)}(\iota(n_2))$ gives the rate of words in the definition of $\iota(n_1)$ that are shared by $\iota(n_2)$. Note that the lack of symmetry in the definition of this index (and also in q_μ^L and q_μ^R) is a natural feature. For example, one can expect that

$$q_\mu^L(\iota(acculturation), \iota(culture)) > q_\mu^L(\iota(culture), \iota(acculturation)), \tag{16}$$

or at least that there is one more element in the counting of the left-hand side, since the word "culture" appears in all the definitions of "acculturation", but the reverse relation could not happen.

A different measure can be used if we want the nouns of N to have different relevance in the model. For example, we can have nouns associated with the central meaning of the set N, and common nouns that are often used in similar contexts and serve to describe the main words. If we come back to the example of the Sociology of Migration, words such as "culture", "acculturation", and "assimilation" are central, and we assign them a weight equal to 1, and words such as "process" or "change", that are used to describe the others, can be weighted by $1/2$. That is, the fact that the word "culture" appears in the definition of "acculturation" is central for the coding of its meaning, while the occurrence of the word "process" just denotes that it is an action. If we divide the words of N in two sets, "Central" and "Complementary", the following measure can be considered instead of the counting measure,

$$\mu(A) = |A \cap Central| + \frac{1}{2}|A \cap Complementary|, \quad A \in \Sigma(N). \tag{17}$$

Although the choice of certain words is arbitrary if we consider them as signs, i.e., taking into account only their denotative character, it is clear that there are relationships given by the linguistic structure of these words that could be taken into account in the relationship between the terms, but that have not been introduced in the model. If we consider the words "culture" and "acculturation", even if a person does not know the meaning of the word "acculturation", that person is unlikely to think that it denotes a new type of carrot that can be found in the local market: he or she is likely to think that it is something related to culture. However, it is difficult to introduce such information a priori in a relational model such as the one we propose. Again, mathematical description is restricted by its own formal limits.

4.2. Word Neighborhoods in a Text

A common criterion to know the intensity of the relationship between words is given by the number of interactions between them, defined as a measure of how two given words appear close to each other in texts written in a given language. To quantify this relationship, we can use the following mechanism: given a natural number $n \geq 2$, two words, w_1 and w_2, in a text are understood to be related by proximity if there are fewer than $n-2$ words between w_1 and w_2. Let us show how this idea can be formalized by means of a set-word embedding, and what the natural distance and the semantic index given by the model mean (Figure 2).

Fix a specific text with N words and define the set S of all the sequences s of size $n \leq N$ of consecutive words in it, indexed by its order of appearance; it can be easily seen that $|S| = N - n + 1$. Fix a specific set of (different) such words $W \subset N$, and write $H(s)$ for the set of words of the sequence s. Define the $\sigma-$algebra of all the subsets of S, and consider

the set-word embedding ι that sends each noun w in W to the set of all the sequences in which w appears, that is, $\iota : W \hookrightarrow \Sigma(S)$,

$$\iota(w) := \{s \in S : w \in H(s)\} \in \Sigma(S), \qquad w \in W,$$

that is, $\iota(w)$ is the set of sequences that contain the word w.

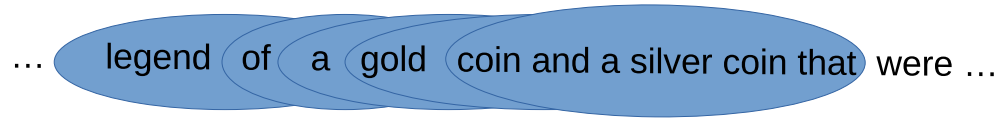

Figure 2. Sequences s in a text for $n = 6$. The last circle marks the sequence s_0, which contains the set $H(s_0) = \{$"*coin*", "*and*", "*a*", "*silver*", "*that*"$\}$, and the previous one, s_1, contains the words $H(s_1) = \{$"*gold*", "*coin*", "*and*", "*a*", "*silver*"$\}$.

Let us show how this construction works for the analysis of words in the text. Fix μ to be the counting measure. Following the definition, the canonical distance q_μ between two words w_1 and w_2 is given by $q_\mu(\iota(w_1), \iota(w_2)) = \max\{q_\mu^L(\iota(w_1), \iota(w_2)), q_\mu^R(\iota(w_1), \iota(w_2))\} = \max\{|\iota(w_1) \cap \iota(w_2)^c|, |\iota(w_1)^c \cap \iota(w_2)|\}$.

Note that $|\iota(w_1) \cap \iota(w_2)^c|$ is the number of sequences of the model that contain w_1 but do not contain w_2, and $|\iota(w_1)^c \cap \iota(w_2)|$ is the same, but changing w_1 by w_2.

The semantic index is given by the expression

$$P_{\iota(w_1)}(\iota(w_2)) = \frac{|\iota(w_1) \cap \iota(w_2)|}{|\iota(w_1)|}, \tag{18}$$

which means the ratio among the number of sequences that contain both w_1 and w_2 and the number of sequences that contain w_1. This clearly gives a natural quantification of the interaction among w_1 and w_2 normalized by the "relevance" of w_1 in the text represented by $|\iota(w_1)|$. In other words, $P_{\iota(w_1)}(\iota(w_2))$ gives a measurement of how far the term w_2 is involved in the function of w_1 as a word in the text.

For example, if both w_1 and w_2 are nouns, $P_{\iota(w_1)}(\iota(w_2))$ would be interpreted as an index of how the meaning of w_1 relates to that of w_2 in the text. Clearly, for $w_1 = w_2$, we obtain $P_{\iota(w_1)}(\iota(w_2)) = 1$, and the index is equal to 0 if all occurrences of the words w_1 and w_2 in the text are separated by more than $n - 2$ words, that is, if every occurrence of w_2 is "outside the textual environment" of any occurrence of w_1.

4.3. Scientific Documents in arXiv

Consider a set, S, of documents related to a certain topic in a scientific preprint repository (e.g., arXiv). Identify each of these documents with the set of all the words that appear in it. Take a set N of scientific terms on this topic that appear in at least one of these papers, and consider the set-word embedding $\iota : n \in N \hookrightarrow \Sigma(N)$, which sends each word n to the set of documents $\iota(n)$ in which it appears, where $\Sigma(S)$ is the $\sigma-$algebra of all the subsets of S.

The question about the terms to be analyzed is to what extent different terms appear together in the documents or, in other words, how close they are in relation to their use in the documents' class. Note that if two terms occur together in all documents, they are inseparable in the word embedding and should be considered a single semantic entity (otherwise q_μ is not a distance).

Take, again, the counting measure. The distance $q_\mu(n_1, n_2)$ indicates the size of the set of documents that either contain n_1 and not n_2, or contain n_2 and not n_1. If $q_\mu(n_1, n_2)$ is small, there are no documents in which the two terms are related, so they are not connected in the semantic environment defined by the document set.

The semantic index is in this case given by

$$P_{\iota(n_1)}(\iota(n_2)) = \frac{\textit{Preprints that contain } n_1 \textit{ and } n_2}{\textit{Preprints that contain } n_1}. \tag{19}$$

It can be easily seen that, unlike in the case of the classical Jaccard index, non-symmetry is a natural feature of relational indexes in set-word embedding models. If n_1 is a rare term in the subject we have fixed to determine the set of preprints, we can expect that it does not appear often in the documents of the fixed collection. But if every time it appears it is accompanied by n_2, we obtain that n_2 is, indeed, relevant to the meaning of n_1 in this context, and then the index takes the value 1. In this case, however, the inverse index $P_{\iota(n_2)}(\iota(n_1))$ might be very small if n_2 is a very common term; that is, n_1 is not relevant for the meaning of n_2.

5. Set-Based Word Embeddings Versus Vector-Valued Word Embeddings: General Set Representation Procedure

The notion of set-word embedding is primarily thought to be as simple as possible. The complexity of the model and all the features that can be added to it are supposed to be performed by means of Lipschitz functions acting on the set metric space. However, it can be easily seen that both constructions are essentially equivalent, although our set-based construction aims to be simpler. In this section, we prove an equivalence result that shows a canonical procedure to pass from a class of models to the other class, and vice versa.

The following result also provides explicit formulae for the transformation. It is, therefore, the main result of the present work, as it allows us to identify any word embedding with a set-word embedding using a canonical procedure. The main advantage of this idea is that the basic information in the set algebra model lies in the measure of the size of the shared meaning between two terms, whereas a standard set-word embedding in a Euclidean space usually occurs automatically, and there is no possibility to interpret how the distances are obtained and what they mean. Thus, set-word embeddings are interpretable, while standard word embeddings are not. We will return to this central point in the Conclusions section of the article.

In addition to its simple formal structure and the advantages of defining word embedding features as pure metric objects (Lipschitz maps, without the need for any kind of linear relation), there is a computational benefit that makes it, in a sense, better. We will explain it after the theorem. As usual, we assume that an isometry between metric spaces is, in particular, bijective.

Theorem 2. *Let W be a (finite) set of word/semantic terms of a certain language. Then, the following statements hold.*

(i) *If $i : W \hookrightarrow X$ is a metric word embedding into a finite metric space (X, d), there is a set-word embedding $\iota : W \hookrightarrow \Sigma(S)$ into a $\sigma-$algebra of subsets $\Sigma(S)$ and a (non-necessarily positive) measure μ on S such that $(i(W), d)$ and $(\iota(W), q_\mu)$ are isometric.*

(ii) *Conversely, if there is a finite set-word embedding $\iota : W \hookrightarrow \Sigma(S)$, there is a metric word embedding $i : W \hookrightarrow X$ into a metric space (X, d) such that $(i(W), d)$ and $(\iota(W), q_\mu)$ are isometric.*

Moreover, every metric word embedding can be considered as normed-space-valued.

Proof. Let us first prove (i). Consider the set of terms W, and the set $i(W) \subset X$. Write $|W| = n$. We can assume that $n > 2$; for $n = 1$ there is nothing to prove, and for $n = 2$ the result is trivially proved by a direct construction with a set, S, of two elements.

Number the elements of W and identify $i(W)$ with the set $\{1, ..., n\}$, which is considered to have the same metric d as $i(W)$. Write $D = (d_{i,j})_{i,j=1}^{n}$, $d_{i,j} = d_{j,i}$ for the metric matrix of d. We want to construct a measure space $(S, \Sigma(S), \mu)$ and a word embedding $\iota : W \hookrightarrow \Sigma(S)$ such that $(i(W), d)$ and $(\Sigma(S), q_\mu)$ are isometric.

Set $S = \{(i,j) : 1 \leq i \leq j \leq n\}$, $\Sigma(S)$ the $\sigma-$algebra of the subsets of S, and consider the word embedding $\iota : W \hookrightarrow \Sigma(S)$ given by

$$\iota(w_i) = \{(j,i) : 1 \leq j \leq i\} \cup \{(i,j) : i < j \leq n\}, \quad i \in \mathbb{N},$$

where $w_i \in W$ is the word with the number i. We have to find a measure μ on S such that for $i \neq j$, $\tau_{i,j} = \mu\big(\iota(w_i) \cup \iota(w_j)\big) - \mu\big(\iota(w_i) \cap \iota(w_j)\big) = d_{i,j} = d_{j,i}$. That is, for $1 \leq i, j \leq n$, $i \neq j$, $\tau_{i,j} = \sum_{k=1,k\neq j}^{i} \mu(\{(k,i)\}) + \sum_{k=i+1,k\neq j}^{n} \mu((i,k)) + \sum_{k=1,k\neq i}^{j} \mu((k,j)) + \sum_{k=j+1,k\neq i}^{n} \mu((j,k)) = d_{i,j}$. Note that $\tau_{i,i} = 0$ for all $1 \leq i \leq n$. Write T for the symmetric matrix $T = (\tau_{i,j})_{i,j}^{n}$, where $\tau_{i,j} = \tau_{j,i}$. Let us write all the equations above using a matrix formula. Consider the matrix $M = (m_{i,j})_{i,j=1}^{n}$, with $m_{i,j} = 1$ if $i \neq j$, and $m_{i,i} = 0$, for all $1 \leq i, j \leq n$.

Write the symmetric matrix N of the coefficients $x_{i,j} = \mu(\{(i,j)\}) = x_{j,i}$, $1 \leq i \leq j \leq n$ that we want to determine. Take any diagonal matrix Δ and define the symmetric matrix $D^* = D + \Delta$. Note that we can write an equation that coincides with $T = D$ for all the elements out of the diagonal as

$$M \cdot N + N \cdot M = D^*,$$

in which the elements of the diagonal take arbitrary values that can be used to normalize the coefficients.

Now we claim that $M^{-1} = \frac{M-(n-2)I_n}{n-1}$, where I_n is the identity matrix of order n. Indeed, note that M^2 has all the elements of the diagonal equal to $n-1$, and all the elements out of the diagonal are equal to $n-2$. Thus,

$$M \cdot (M - (n-2)I_n) = (M - (n-2)I_n) \cdot M = M^2 - (n-2)M = (n-1)I_n.$$

Now, consider the equations

$$M^{-1}MNM^{-1} + M^{-1}NMM^{-1} = NM^{-1} + M^{-1}N = M^{-1}D^*M^{-1},$$

that give $N \cdot (M - (n-2)I_n) + (M - (n-2)I_n) \cdot N = (n-1)M^{-1}D^*M^{-1}$. Then, $NM - (n-2)N + MN - (n-2)N = (n-1)M^{-1}D^*M^{-1}$, and so, using that $MN + NM = D^*$, we obtain the symmetric matrix

$$N = \frac{1}{2(n-2)}\big(D^* - (n-1)M^{-1}D^*M^{-1}\big).$$

This gives a result that is not unique, as it depends on the diagonal matrix Δ. Note that, due to the required additivity of the measure μ, the set of all the values $\mu(\{(i,j)\})$ determine a measure μ, which is not necessarily positive.

The proof of (ii) is obvious; take a finite measure space $(S, \Sigma(S), \mu)$ and a set-word embedding $\iota : W \hookrightarrow \Sigma(S)$, $|W| = n$. Suppose that we have a numbering in W. By definition, the set $X = \{x_1 = \iota(w_1), ..., x_n = \iota(w_n)\} \subseteq \Sigma(S)$ is a metric space with the distance q_μ, so the map $w_i \mapsto x_i \in (X, q_\mu)$ is the required word embedding.

We only need to prove that we can assume that (X, q_μ) is a subset of a normed space $(E, \|\cdot\|_E)$, and so we can define the word embedding ι to have values into $(E, \|\cdot\|_E)$. There are a lot of different representations that can be used; probably the simplest one is given by

the identification of the metric space (X, q_μ) with the Arens–Eells space $(AE(X), \|\cdot\|_{AE})$, which is explained in the Introduction. It is well known (see, for example, [26], Ch.1) that there is an isometric Lipschitz inclusion $h : (X, q_\mu) \hookrightarrow (AE(X), \|\cdot\|_{AE})$ given by $h(x) = m_{x,0}$, where 0 is a (could be arbitrarily chosen) distinguished element of X, and $m_{x,0}$ is the molecule defined by x and 0. Therefore, $h \circ \iota : W \hookrightarrow AE(X)$ is the desired vector word embedding. Of course, given that $h \circ \iota(W)$ is a finite set in the Banach space $(AE(X), \|\cdot\|_{AE})$, we can represent it with coordinates to obtain a vector-like representation such as the reader might identify with a usual vector word embedding. This finishes the proof. □

Remark 2. *The proof of Theorem 2 gives useful information, which is, in fact, its main contribution for the representation of word embeddings by means of subsets. It gives an explicit formula to compute, given a vector-valued word embedding i, a measure space whose standard metric space structure provides an isometric representation to the one given by i. Indeed, for $n \geq 2$, it is given by the formula*

$$N = \frac{1}{2(n-2)}\Big((D+\Delta) - (n-1)M^{-1}(D+\Delta)M^{-1}\Big),$$

where Δ is a diagonal matrix with free parameters, $M = \begin{bmatrix} 0 & 1 & ... & 1 \\ 1 & 0 & ... & 1 \\ ... & ... & ... & ... \\ 1 & 1 & ... & 0 \end{bmatrix}$, and

$$M^{-1} = \frac{M-(n-2)I_n}{n-1} = \frac{1}{n-1}\begin{bmatrix} 2-n & 1 & ... & 1 \\ 1 & 2-n & ... & 1 \\ ... & ... & ... & ... \\ 1 & 1 & ... & 2-n \end{bmatrix}.$$

This equation gives the values of the measure μ for the atoms of the class

$$\{(1,1),(1,2),...,(i,j),...,(n,n) : 1 \leq i \leq j \leq n\}$$

that are represented in the symmetric matrix N. They can be used to compute the values of the distances between the elements of the representation,

$$\iota(w_i) = \{(k,i) : 1 \leq k \leq i\} \cup \{(i,k) : i+1 \leq k \leq n\}, \quad i = 1,...,n,$$

which are given by

$$q_\mu(\iota(w_i), \iota(w_j)) = \mu(\iota(w_i) \cup \iota(w_r)) - \mu(\iota(w_i) \cap \iota(w_r)).$$

In the model, the diagonal matrix Δ and the metric q_μ, are key elements that shape its behavior. Adjusting the diagonal entries of Δ—which work as parameters of the model—tunes the metric, modifying the relative weights of dimensions to meet specific constraints or properties. The metric q_μ defines the distance measure, guiding how distances between terms are evaluated by means of their representation as classes of subsets. Together, Δ and the construction of q_μ allow the model to adapt flexibly to problem-specific requirements, ensuring robustness and alignment with desired outcomes.

Example 2. *Take $n = 3$, and consider the word embedding $i : W \hookrightarrow W$, where $W = \{w_1, w_2, w_3\}$. These terms can be, for example, three nouns in a model that we want to develop, as $\{man, king, queen\}$. The distance matrix suggested below is coherent with the idea that "man" is, in a sense, close to "king" (a king is a man), and "king" is close to "queen" (both belong to royalty), but, on a comparative scale, "man" and "queen" are not so close.*

- *Endow W with the metric d given by the matrix $D = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}$. We proceed as in the proof of Theorem 2; obtain the set*

$$X = \{(1,1),(1,2),(1,3),(2,2),(2,3),(3,3)\}.$$

We consider the set-word embedding $\iota : W \hookrightarrow \Sigma(X)$ defined as in the proof of the theorem; that is, for instance, $\iota(w_1) = \{(1,1),(1,2),(1,3)\}$ (below for the other terms). This means, for example, that the word "king" is represented by a set of three (vector) indices, $\{(1,1),(1,2),(1,3)\}$, and so on. These indexes could represent different characteristics of the semantic term but, from a formal point of view, they are only distinguished indexes. Take the free parameter diagonal matrix as $\Delta = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$. Then,

$$N = \frac{1}{2}\left(\begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix} - 2\begin{bmatrix} -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{bmatrix}\begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}\begin{bmatrix} -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{bmatrix}\right)$$

$$= \frac{1}{2}\left(\begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix} - 2\begin{bmatrix} -1 & 1/2 & 1 \\ 1/2 & 0 & 1/2 \\ 1 & 1/2 & -1 \end{bmatrix}\right) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

This matrix gives the values of the measure μ for the elements of the atoms of the measure space: $\mu(\{(1,1)\} = 1$, $\mu(\{(1,2\}) = 0$, $\mu(\{(1,3)\}) = 0$, and so on. In the model, they give the measure (understood as weight) of the different characteristics that represent each of the atomic indices (k,j). Using that to verify that the calculations coincide with the values of the original distance is straightforward. For example,

$$q_\mu(\iota(w_1),\iota(w_2)) = \mu(\{(1,1),(1,2),(1,3),(2,2),(2,3)\}) - \mu(\{(1,2)\})$$

$$= \mu(\{(1,1)\}) + \mu(\{(1,3)\}) + \mu(\{(2,2)\}) + \mu(\{(2,3)\})$$

$$= 1 + 0 + 0 + 0 = 1 = d(i(w_1), i(w_2)),$$

while

$$q_\mu(\iota(w_1),\iota(w_3)) = \mu(\{(1,1),(1,2),(1,3),(2,3),(3,3)\}) - \mu(\{(1,3)\})$$

$$= \mu(\{(1,1)\}) + \mu(\{(1,2)\}) + \mu(\{(2,3)\}) + \mu(\{(3,3)\})$$

$$= 1 + 0 + 0 + 1 = 2 = d(i(w_1), i(w_3))$$

which, of course, coincide with the corresponding coefficients of the original metric matrix D. Note that the measure μ is equal to 0 for most of the atoms, so it cannot distinguish between certain non-empty sets of the generated $\sigma-$algebra. However, the formula of our distance defined using the set algebra separates between all the elements of the canonical set-word embedding $\iota : \{w_1, w_2, w_3\} \hookrightarrow \Sigma(X)$, given in this case, as we said, by (see Figure 3)

$$\iota(w_1) = \{(1,1),(1,2),(1,3)\}, \quad \iota(w_2) = \{(1,2),(2,2),(2,3)\},$$

and $\iota(w_3) = \{(1,3),(2,3),(3,3)\}.$

The subsets represented in Figure 3 can be understood as subsets of different features (the indices (k, j)) that allow the representation of the original semantic environment $\{man, king, queen\}$, in the sense that each subset of features represents different semantic objects with their own semantic role in the model.

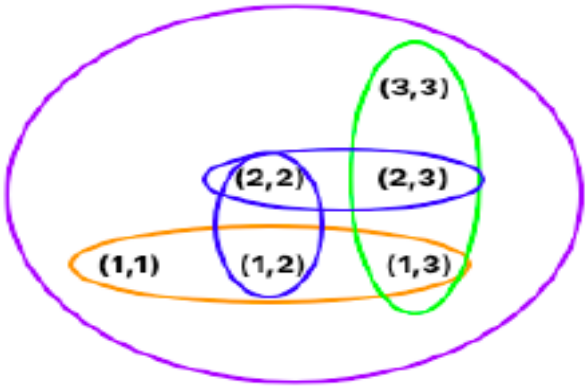

Figure 3. Sets defining the representation of the set-word embedding.

- *Let us show now the same computations for a different metric matrix D such that all the distances between the three points are different. In this case, we obtain*

$$N = \frac{1}{2}\left(\begin{bmatrix}0&1&2\\1&0&3\\2&3&0\end{bmatrix} - 2\begin{bmatrix}-\frac{1}{2}&\frac{1}{2}&\frac{1}{2}\\\frac{1}{2}&-\frac{1}{2}&\frac{1}{2}\\\frac{1}{2}&\frac{1}{2}&-\frac{1}{2}\end{bmatrix}\begin{bmatrix}0&1&2\\1&0&3\\2&3&0\end{bmatrix}\begin{bmatrix}-\frac{1}{2}&\frac{1}{2}&\frac{1}{2}\\\frac{1}{2}&-\frac{1}{2}&\frac{1}{2}\\\frac{1}{2}&\frac{1}{2}&-\frac{1}{2}\end{bmatrix}\right)$$

$$= \frac{1}{2}\left(\begin{bmatrix}0&1&2\\1&0&3\\2&3&0\end{bmatrix} - 2\begin{bmatrix}0&1/2&1\\1/2&-1&3/2\\1&3/2&2\end{bmatrix}\right) = \begin{bmatrix}0&0&0\\0&1&0\\0&0&2\end{bmatrix}.$$

Example 3. *Now take $n = 4$ and $W = \{w_1, w_2, w_3, w_4\}$. Then, we have that, for $D = \begin{bmatrix}0&1&2&2\\1&0&3&2\\2&3&0&2\\2&2&2&0\end{bmatrix}$, the same computations as in the other examples give the following matrix defined by the values of the measure μ for the atoms of the representation, which are, in this case,*

$$S = \{(1,1),(1,2),(1,3),(1,4),(2,2),(2,3),(2,4),(3,3),(3,4),(4,4)\}.$$

The measure matrix N is

$$\frac{1}{4}\begin{bmatrix}0&1&2&2\\1&0&3&2\\2&3&0&2\\2&2&2&0\end{bmatrix} - \frac{1}{4}\begin{bmatrix}-2&1&1&1\\1&-2&1&1\\1&1&-2&1\\1&1&1&-2\end{bmatrix}\begin{bmatrix}0&1&2&2\\1&0&3&2\\2&3&0&2\\2&2&2&0\end{bmatrix}\begin{bmatrix}-\frac{2}{3}&\frac{1}{3}&\frac{1}{3}&\frac{1}{3}\\\frac{1}{3}&-\frac{2}{3}&\frac{1}{3}&\frac{1}{3}\\\frac{1}{3}&\frac{1}{3}&-\frac{2}{3}&\frac{1}{3}\\\frac{1}{3}&\frac{1}{3}&\frac{1}{3}&-\frac{2}{3}\end{bmatrix}$$

$$= \frac{1}{4}\left(\begin{bmatrix}0&1&2&2\\1&0&3&2\\2&3&0&2\\2&2&2&0\end{bmatrix} - \begin{bmatrix}-2&0&2&3\\0&-4&4&2\\2&4&-6&1\\3&2&1&-4\end{bmatrix}\right) = \begin{bmatrix}1/2&1/4&0&-1/4\\1/4&1&-1/4&0\\0&-1/4&3/2&1/4\\-1/4&0&1/4&1\end{bmatrix}.$$

For example,

$$q_\mu(\iota(w_2), \iota(w_3)) = \mu(\iota(w_2) \cup \iota(w_3)) - \mu(\iota(w_2) \cap \iota(w_3))$$

$$= \mu(\{(1,2),(2,2),(2,3),(2,4),(1,3),(3.3),(3,4)\} - \mu(\{(2,3)\})$$
$$= \mu(\{(1,2)\}) + \mu(\{(2,2)\}) + \mu(\{(2,4)\}) + \mu(\{(1,3)\}) + \mu(\{(3.3)\}) + \mu(\{(3.4)\})$$
$$= 1/4 + 1 + 0 + 0 + 3/2 + 1/4 = 3,$$

which coincides with the coefficient of D in the position $(2,3)$*, that is,* $d(i(w_2), i(w_3))$*. Note that, according to the measure matrix N, the measure* μ *is not positive: there are atoms for which the measure is negative, and others for which the measure equals* 0*. However, the standard formula for the associated set distance gives a proper distance matrix, as it coincides with the metric matrix D.*

It should be noted that we need the most abstract notion of set-word embedding to obtain that, in general, any (metric) word embedding can be written as a set-word embedding: the measure μ obtained for the representation is not necessarily positive, but it has to be positive for all the sets defined as symmetric differences of the elements of the $\sigma-$algebra in the range of the representation $\iota : W \hookrightarrow \Sigma(S)$, that is, for the elements such as $(\iota(w_i) \cup \iota(w_j)) \setminus (\iota(w_i) \cap \iota(w_j))$, for which we need positive measure evaluations to obtain a suitable metric.

Note also that the standard application of the procedure, provided by the equations explained above, lies in the identification of semantic terms with elements of a $\sigma-$algebra of subsets. The size of such subset structures increases exponentially with the number of semantic terms, which could compromise the scalability of the method. This means that it might be necessary to imagine a representation procedure using subsets for each specific problem, which would detract from the generality of our technique.

Remark 3. *Direct applications of word embeddings, such as word similarity and analogy completion, should be approached from a different point of view than in the case of vector-valued word embeddings. In the case of word similarity, the linear space underlying the representation facilitates the task, as the scalar product inherent in Euclidean space provides the correlation as well as the cosine. In our case, however, a different approach is necessary, since, a priori, there is no associated linear space other than the one given by our representation theorem, which is often too abstract, as we just said. But it is possible to obtain a quantification of the notion of word similarity by relating to each word the calculated semantic index with respect to the other semantic terms of the model, thus constructing a vector whose correlation with any other vector associated with another word can already be calculated. Again, the advantage of our method compared to embedding words in a Euclidean space is interpretability: each coordinate of the vector thus constructed represents the extent to which the word and any other term in the model share their meaning, in the sense indicated by the semantic index used. Analogy completion could also be better performed in the case of embeddings of sets of words, as the representation of the word and any other term in the model share the same meaning, in the sense indicated by the semantic index used. Analogy completion could also be better realized in the case of word-set embeddings, since the representation is based on the extent to which two terms share their meaning, and then the logical relationships are supported by the mathematical formalism.*

6. Discussion: Comparing Multiple Word Embeddings with the Set-Word Embedding in an Applied Context

In this section, we use the tools developed through the paper to obtain a set-word embedding related to the semantic indices (semantic projections) that are explained in Section 3 (Definition 2). We follow the model presented in Section 4.3 to define a word embbeding associating with each term a class of subsets of documents in which the word appears. To facilitate reproducibility for the reader, we opted to use the Google search engine for the calculations rather than relying on a scientific document repository, which

may have access restrictions. For this application, we utilize the results provided by Google searches, meaning that the measure of the set representing a given term corresponds to the number of documents indexed by Google that contain the word in question.

Let us fix the term "gold" and consider the set of words given by the Word2vec embedding (Figure 4). As usual, the associated vector space is endowed with the Euclidean norm. To compare our word-set embedding with this one, we use as working environment the 10 words that are close with respect to the distance provided by the embedding to the fixed term, including "gold". The terms are {*gold*, *silver*, *medal*, *blue*, *reserves*, *coin*, *rubber*, *diamonds*, *tin*, *timber*}, and the distances to the term "gold" can be found in the second column of Table 1.

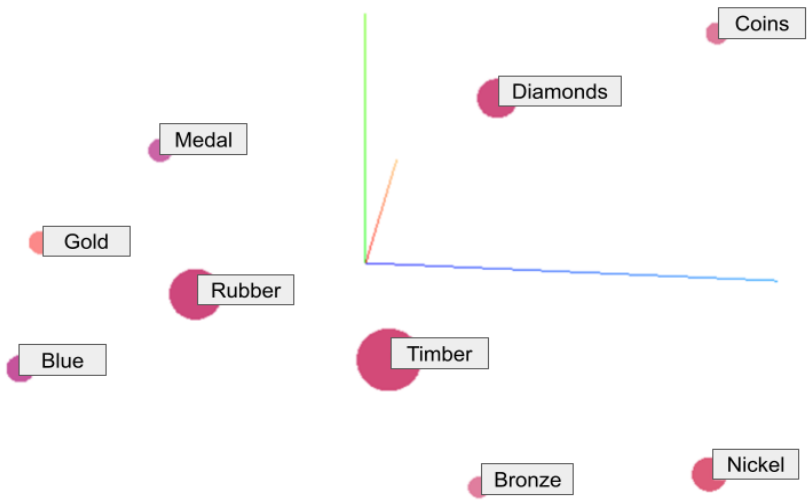

Figure 4. A 3D representation of the embedding of the closest words to the term "gold", according to the Euclidean distance, using Word2vec Google News (71291x200) (http://projector.tensorflow.org/ (accessed on 3 January 2025)).

The other values of Table 1 complete the information to compute the metric q_μ provided by the set-word embedding using Google search. Following the theoretical development explained in Section 3, the basic elements to understand our model are the semantic indices P_A and P_B, which are shown in Figures 5 and 6, respectively.

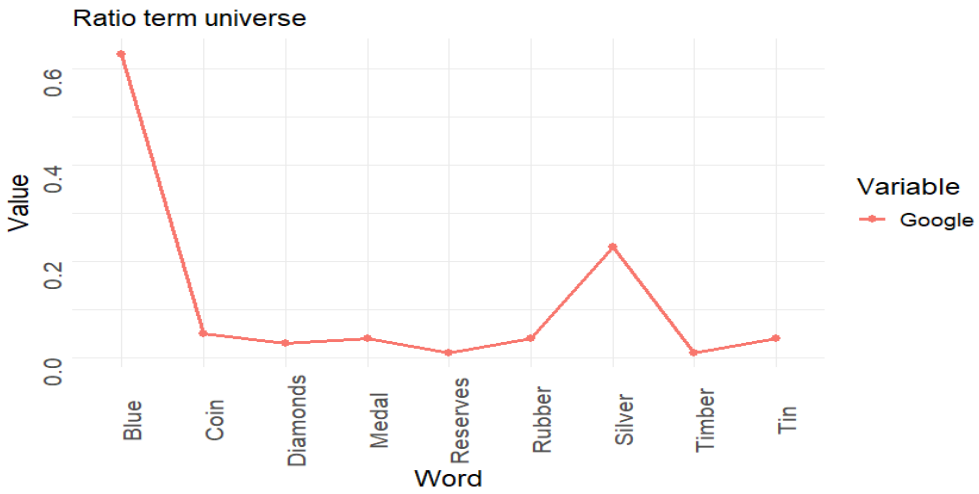

Figure 5. Semantic indices P_A for A =gold ($q = q_\mu$).

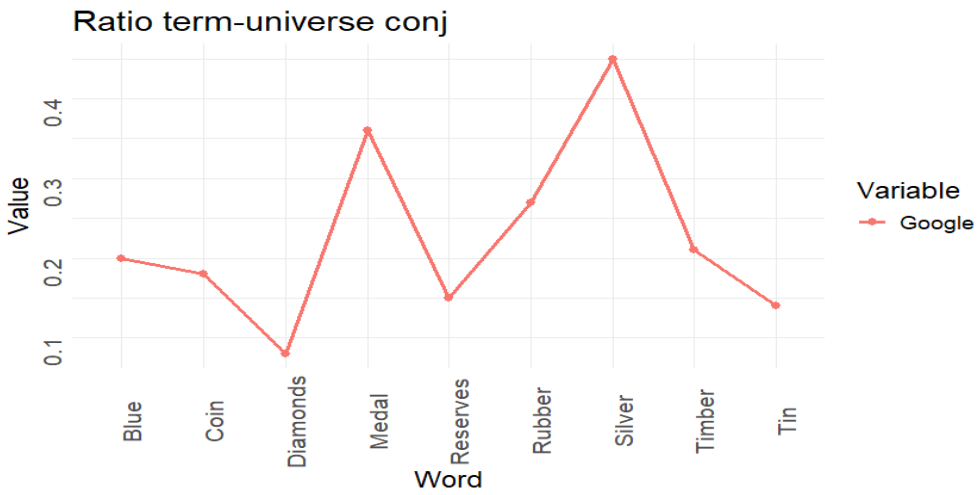

Figure 6. Conjugate semantic indices P_B for B for each of the selected words.

Recall that we have $q_\mu = q_\mu^L + q_\mu^R$, and using the formulas

$$q_\mu^L(A,B) = \mu(A)\Big(1 - P_A(B)\Big) \quad \text{and} \quad q_\mu^R(A,B) = \mu(B)\Big(1 - P_B(A)\Big)$$

given by Lemma 1, we can compute Table 1. In this table, the measures μ are written in billions (1,000,000,000). To simplify calculations, we will divide by this number in the distance definition computed below, reducing unnecessary complexity.

Table 1. Table with the words, the values of the Euclidean distance with the word "gold", and the corresponding values of the elements of the set-word embedding.

Word	W2vec	P_A	P_B	$\mu(A)$	$1 - P_A$	$\mu(B)$	$1 - P_B$	q_μ
silver	0.786	0.233	0.476	7.840	0.767	4.110	0.524	8.166
medal	1.129	0.03	0.362	7.840	0.97	573	0.638	7.97
blue	1.141	0.64	0.204	7.840	0.36	25.270	0.796	22.937
reserves	1.135	0.01	0.149	7.840	0.99	542	0.851	8.222
coin	1.136	0.046	0.185	7.840	0.954	1.460	0.815	8.670
rubber	1.456	0.027	0.27	7.840	0.973	771	0.73	8.191
diamonds	1.109	0.021	0.088	7.840	0.979	379	0.912	8.021
tin	1.097	0.032	0.136	7.840	0.968	1.820	0.864	9.161
timber	1.125	0.02	0.21	7.840	0.98	288	0.79	8.004

To facilitate the comparison of the metrics, in Figures 7 and 8 we divide the distance associated with the set-word embedding by eight. Obviously, this change of scale does not affect the properties relevant to the comparison. As can be seen, the distributions obtained for the two metrics are similar, although there are significant differences. Figure 8 shows the same information as Figure 7, but in it we have removed the term "blue" to facilitate the comparison of the values of the rest of the terms, as its large values disturb the overall view.

The primary reason for these differences is the influence of context. Traditional early word embeddings rely on a fixed, universally applicable mapping of semantic distances, while the set-word embedding allows the incorporation of contextual information derived from the large variety of documents indexed by Google. Although the boundaries of this contextual information are not entirely clear, it is evident that the relationships between semantic terms are influenced by the volume of information available across internet documents. This explains, for instance, why the term "blue" is quite distant from "gold",

as the various meanings of "blue" are not statistically related to the meaning of "gold" in a significant proportion.

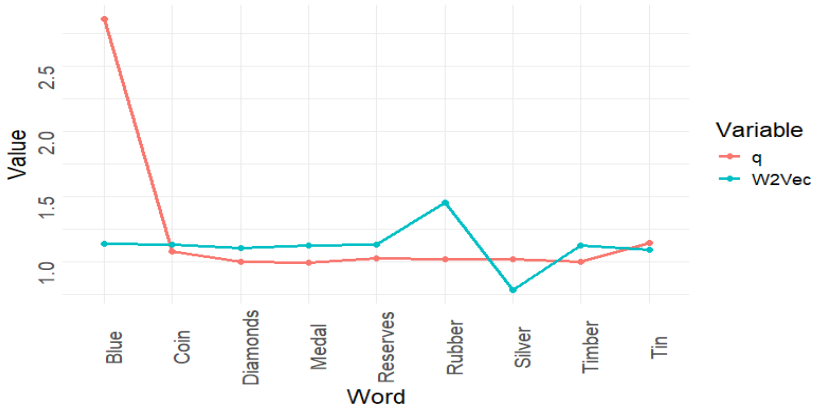

Figure 7. Comparison of the results provided by Word2vec and the set-word embedding defined by Google search for the term "gold".

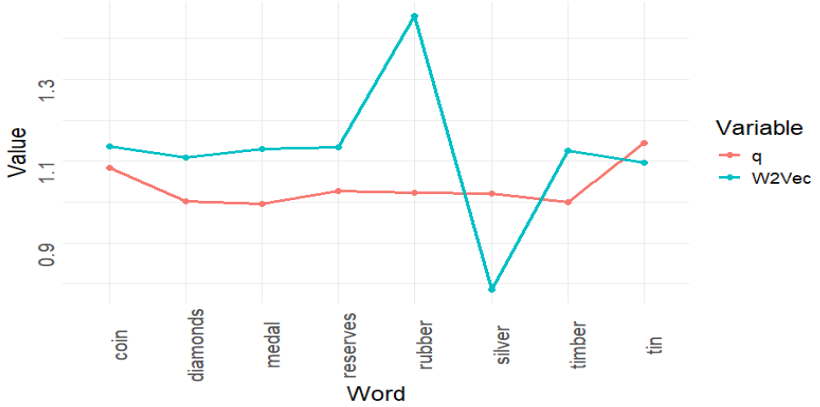

Figure 8. Comparison of the results provided by Word2vec and the set-word embedding excluding the term "blue" for a better visualization.

The term "medal" is closer to "gold" than to "silver" in the information repository queried by Google. This makes sense when we consider that "gold medal" is a much more prevalent phrase on the internet, highlighting common usage patterns. In contrast, while gold and silver are similar in their inherent nature as metals, their association in language is less frequent compared to the popularized use of "gold medal". Hence, usage, rather than inherent similarity, drives this result. The same can be argued regarding the terms "reserves", "rubber", or "coin". The relationships provided by the word-set embedding with the other words, "diamonds", "tin", and "wood", are more conventional, as can be seen by comparing them with the results provided by Word2vec.

6.1. Advanced Models

Let us introduce now in the discussion other word embeddings that have been designed following different ideas. We now compare some advanced word embedding models, focusing on their ability to capture semantic relationships between precious metals, other commodities, and related terms. The evaluation includes performance metrics and similarity analysis using cosine similarity and L2 distance. As this information is quite exhaustive, we preferred to move some of it to Appendix A.

Microsoft Research's E5 models, including E5-small (384 dimensions) and E5-base (768 dimensions), represent significant advances. Trained by contrastive learning on diverse datasets, they are versatile in their applications. We also examined Microsoft's

MiniLM models, which offer effective alternatives through knowledge distillation. The L6 variant emphasizes speed, while L12-v2 offers deeper semantic understanding. Finally, ByteDance's BGE-small model, optimized for retrieval tasks, combines contrastive learning and masked language modeling to deliver high performance in a compact form factor. These models offer different trade-offs between efficiency and semantic accuracy. More information can be found in Appendix A (see the bibliographic references to related papers in this appendix).

Table 2 provides the results of the distances of the term "gold" to the other terms using these word embeddings, which can be compared with the results given by our set-word embedding and by Word2Vec. Figure 9 gives a clear picture of how these other embeddings behave in comparison with the previous ones. In order to compare in a better way, all the models are normalized to have an average value of about eight.

Table 2. L2 distances of the term "gold" to the other terms using the word embeddings E5-small L2 distance, E5-base L2 distance, MiniLM-L6 L2 distance, BGE-small L2 distance, and MiniLM-L12-v2 L2 distance.

Word Embed.	Silver	Medal	Blue	Reser.	Coin	Rubber	Diam.	Tin	Timber
E5-small	0.499	0.487	0.652	0.620	0.540	0.634	0.552	0.656	0.685
E5-base	3.262	2.804	4.327	2.741	2.817	3.559	2.243	3.296	3.020
MiniLM-L6	0.786	0.849	1.026	1.009	0.838	1.112	1.031	1.217	1.225
BGE-small	1.076	1.244	1.425	1.475	1.199	1.363	1.156	1.369	1.449
MiniLML12v2	1.693	2.328	2.366	2.641	2.047	3.303	2.399	2.208	2.879

Figure 9 presents the results of the seven models included in the comparison, with the set-word embedding highlighted in purple. To facilitate a clearer comparison, we scaled the results by appropriate factors to ensure that the representations are centered. In Figure 10, we show only two selected models alongside the set-word embedding for a more focused analysis. As observed in both figures, the set-word embedding produces smoother results overall (the example is chosen for this to happen), although it still follows the primary trends exhibited by the other models, albeit with some slight deviations.

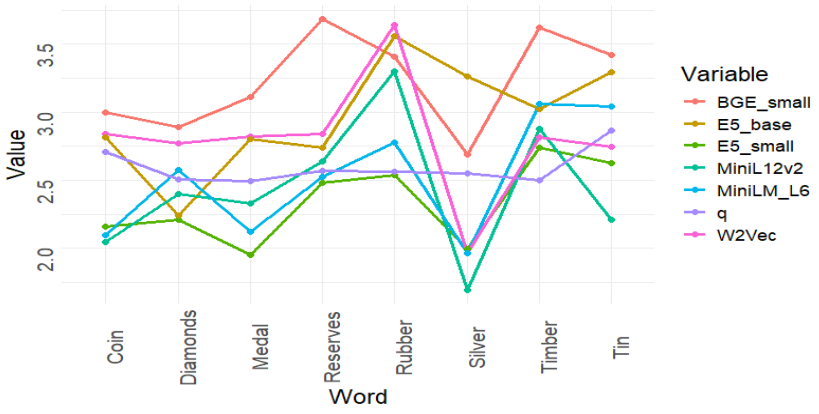

Figure 9. Comparison of the results provided by Word2vec and the set-word embedding including the other word embeddings of Table 2.

For the three models selected for Figure 10 (set-word embedding, E5-small, and MiniLM-L6), the trends are more clearly visible, with the set-word embedding shown in purple. This allows a more focused comparison among these specific models, highlighting the distinctive patterns in their performance.

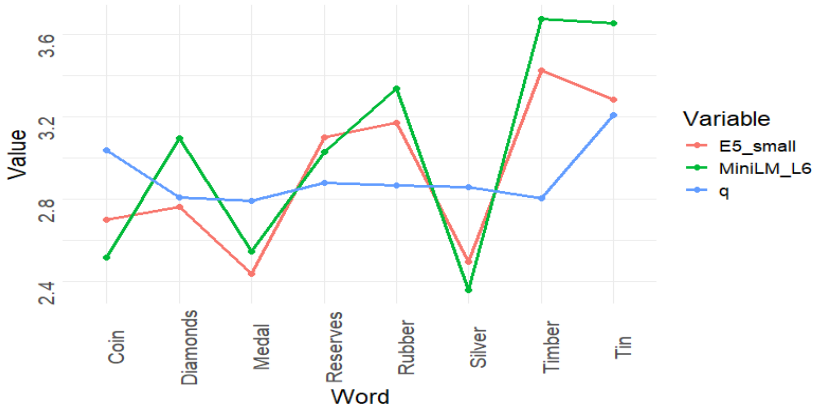

Figure 10. Comparison of the results provided by the set-word embedding and the models E5-small and MiniLM-L6, which are the ones that have a similar distribution.

As can be seen, the comparison of the models in this section reveals that, while they all produce different outputs, the variations are not substantial, and some common trends emerge, especially after adjusting the scale for better visualization. The set-word embedding provides a more stable result compared to the other models. However, when considering the combined contributions of the other models, the set-word embedding generally aligns with the overall averages of their performance.

6.2. Remarks on the Comparison of Semantic Models

In summary, the results of this analysis suggest that if the meanings of words are derived from their relationships within a given semantic environment, those meanings fundamentally depend on the measurement tool used to establish these connections. In fact, it seems more practical for many applications to assume that no isolated relationships between terms fully define a word's meaning. These relationships are always shaped by context, and—this being the main theoretical contribution of this paper—by the measurement tool used to uncover these connections.

Let us give some hints about possible future applications of our ideas. Let us provide some observations on how other commonly used NLP tools can be adapted to our formal context. Techniques such as prompt engineering and few-shot learning, which are fundamental to improve the performance of linguistic models, could be effectively integrated with set-word embeddings. Prompt engineering [30–32] involves crafting specific instructions or queries to guide models like GPT toward generating precise and contextually relevant responses. In our context, the construction of these prompts could be informed by the semantic indices used in set-word embeddings, with logical relationships between terms directly translating into formal properties of the model, thereby simplifying the prompt creation process. A similar approach applies to few-shot learning [33,34], which enables models to perform tasks with minimal examples by leveraging their generalization capabilities within the context of the provided prompt.

Additionally, set-word embeddings can serve as a foundation for developing new procedures for LLM-based semantic search (see, for example, ref. [35] and the references therein). This advanced search methodology employs large language models (LLMs) to move beyond traditional keyword matching, identifying implicit relationships and deeper meanings in text. By incorporating set-word embeddings as the underlying semantic framework, this approach could achieve greater interpretability and semantic precision compared to other word embedding options. The integration of these interpretable embeddings would enhance the relevance and contextual accuracy of retrieved information, offering a more robust and insightful framework for semantic search.

Regarding the limitations of the proposed set-word embeddings, we have shown that our procedure is general from a theoretical point of view, in the sense that it covers all situations where standard vector word embeddings are applied (Theorem 2). However, computing the equations to relate both models could be computationally expensive. As can be seen, the size of the matrices involved in the computations could be huge in real cases, and the number of computations could increase exponentially, since in the basic elements of the model are subsets of the initial set of indices, thus increasing with the power 2^N of the original set of terms N. This problem could also carry over to the usual applications of our model, as the construction of the power set underlying the representation could break the potential scalability of the procedure. Other methods of identifying semantic terms with elements of a $\sigma-$algebra of subsets would have to be used, adapting them to specific contexts, as shown in the examples presented in Section 4. This could restrict their use, as a concrete representation would have to be invented for each application. The development of alternative systematic approaches for defining set-word embeddings depending on the context, along with their analysis and comparison with existing word embedding techniques, represents a key focus of our future work.

7. Conclusions

Set-word embeddings represent a more general approach—though fundamentally distinct from standard methods—from the outset in defining word embeddings on structured spaces. Rather than relying on a linear space, which can sometimes introduce confusion in the representation, the set-word embedding associates each term with a subset within a class of subsets, S, where the class itself has some structural properties. This set-based approach offers a more flexible and intuitive framework for capturing semantic relationships. Furthermore, this original set can always be embedded within a more robust set structure, such as the topology generated by S. This allows the application of topological tools to better understand the embedding process. In this paper, we illustrated the case where S is embedded in the $\sigma-$algebra $\Sigma(S)$ generated by S itself. In this context, a canonical structure can be established, first treating it as a measure space $(S, \Sigma(S), \mu)$, and subsequently as a metric space. The embedding representation, then, is viewed as an embedding in a measure space, offering a new perspective on word embeddings that allows for richer and more flexible analysis.

This set-based embedding framework not only enables a deeper understanding of the structural properties underlying word representations but also provides a means to apply advanced mathematical techniques, such as measure theory and topology, to the problem. By leveraging these tools, we can gain insights into the continuity, convergence, and general behavior of word embeddings in a more formalized and rigorous way. This approach also opens the door to the exploration of new types of embeddings, which could potentially capture more complex relationships between words and their meanings.

The main limitation of the proposed technique lies in its practical performance, particularly in terms of scalability. Enhancing the algorithms for real-time applications may be challenging. However, there is an advantage over other methods: the computation of distances between two semantic terms can be performed independently of the other coefficients in the metric matrix. This is because the semantic indices are defined using external information sources that are specific to each pair of terms. Therefore, if only a small subset of distances is needed, the method could remain competitive, even in the context of large and demanding information resources.

Finally, it is worth noting that this methodology offers a flexible way of defining word embeddings and introduces a shift in the way we understand them. As demonstrated in the example in Section 6, the widely accepted assumption in NLP that context shapes

word relationships can be taken a step further from our perspective. Not only does context influence the relational meaning of words, but the measurement tool used to capture these relationships also defines a specific way of interpreting them. Set-word embeddings provide access to a variety of these measurement tools: for instance, each document repository creates a relational structure that reflects how terms interact within that particular context. The distance q_μ, determined by a given measure μ on the set-word embedding, thus defines a unique interpretation of the words' meanings.

Let us conclude the article by mentioning what is possibly one of the main benefits of the proposed set-word embeddings. Unlike the vast majority of word embeddings, the set-word embedding method offers clear advantages in terms of interpretability and explainability compared to contemporary methods. Although modern language models often rely on post hoc explanation methods such as feature importance, saliency maps, concept attribution, prototypes, and counterfactuals [36], the set-based approach provides inherent interpretability through its fundamentally distinct mathematical foundation, where each term is associated with a subset within a class of subsets, S, which can be embedded in a measure space $(S, \Sigma(S), \mu)$, and subsequently as a metric space. This transparent mathematical structure addresses the lack of formalism in terms of problem formulation and clear and unambiguous definitions identified as a key challenge in XAI research [37]. The method's representation through struc- tured spaces rather than traditional linear space, which can sometimes introduce confusion, provides a clearer framework for semantic analysis, contrasting with current embedding approaches that face faithfulness issues and often require complex post hoc explanations [38]. Furthermore, as noted in this article, the topology generated by S allows the application of topological tools to better un- derstand the embedding process, offering a level of theoretical interpretability that aligns with the call for more attention to interpreting ML models [37]. This mathematical rigor addresses the need for faithful-by-construction approaches to model interpretability, though with more solid theoretical foundations than some current self-explanatory models that fall short due to obstacles like label leakage [38].

Author Contributions: Conceptualization, P.F.d.C. and E.A.S.P.; methodology, E.A.S.P.; software, C.A.R.P.; validation, C.S.A.; formal analysis, C.A.R.P. and E.A.S.P.; investigation, P.F.d.C., C.A.R.P. and E.A.S.P.; data curation, C.S.A.; writing—original draft preparation, E.A.S.P. and C.A.R.P.; writing—review and editing, P.F.d.C.; visualization, C.S.A.; supervision, E.A.S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Generalitat Valenciana (Spain), grant number PROMETEO CIPROM/2023/32.

Data Availability Statement: Data is contained within the article.

Acknowledgments: We would like to acknowledge the support of Instituto Universitario de Matemática Pura y Aplicada and Universitat Politècnica de València.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

This paper presents selected examples of state-of-the-art word embedding models, focusing on their ability to capture semantic relationships between precious metals, commodities, and related terms. The evaluation includes both performance metrics and a detailed similarity analysis using cosine similarity and L2 distance measures.

Let us give first an overview of the models we present in the Discussion section of the paper. The E5 family of models, developed by Microsoft Research [39], represents a significant advancement in text embeddings. We evaluate both E5-small (384 dimensions) and

E5-base (768 dimensions). These models were trained using contrastive learning on diverse datasets including web content, academic papers, and domain-specific documentation. MiniLM models (L6 and L12-v2) are lightweight alternatives developed by Microsoft [40]. Using knowledge distillation techniques, they compress BERT-like architectures while maintaining competitive performance. The L6 variant emphasizes efficiency, while L12-v2 provides more nuanced semantic understanding. BGE-small, developed by ByteDance [41], is optimized for efficient retrieval tasks. It employs a combined training strategy of contrastive learning and masked language modeling, achieving strong performance despite its compact architecture. The results are shown in the following subsections.

Appendix A.1. Performance Metrics

Table A1 shows the computational performance of each model. The metrics include model loading time, average embedding time per input, and embedding dimensionality.

Table A1. Model performance comparison.

Model	Load Time (ms)	Avg Embed Time (ms)	Dimension
E5-small	3015.22	177.77	384
E5-base	4201.97	255.57	768
MiniLM-L6	1688.44	112.07	384
BGE-small	2864.15	233.54	384
MiniLM-L12-v2	4450.30	213.06	384

Regarding the performance characteristics of the models, it can be said that MiniLM-L6 demonstrates superior efficiency with the lowest load and embedding times. On the other hand, E5-base requires significant computational resources but provides higher dimensionality. MiniLM-L12-v2 shows unexpectedly high load time despite having similar architecture to L6, while BGE-small maintains balanced performance metrics.

Appendix A.2. Similarity Analysis

We analyze both cosine similarity and L2 distance matrices for the ten key terms involved in the example that we follow in Section 6: Gold, Silver, Medal, Blue, Reserves, Coin, Rubber, Diamonds, Tin, and Timber. For cosine similarity, higher values (closer to 1.0) indicate greater similarity. For L2 distance, lower values indicate greater similarity.

Appendix A.2.1. E5-Small Results

The E5-small model shows strong relationships between precious metals and related terms (see cosine similarity and L2 distance in Table A2 and Table A3, respectively). Gold–Silver–Medal form a highly cohesive cluster (cosine > 0.938), and precious metals show a strong correlation (Gold–Silver: 0.951). Consistently low L2 distances between related terms are obtained, and Timber shows relatively weaker relationships across all terms.

Table A2. E5-small cosine similarity matrix.

	Gold	Silver	Medal	Blue	Res.	Coin	Rubber	Diam.	Tin	Timber
Gold	1.000	0.951	0.952	0.916	0.925	0.942	0.922	0.941	0.915	0.908
Silver	0.951	1.000	0.938	0.928	0.922	0.936	0.911	0.928	0.932	0.899
Medal	0.952	0.938	1.000	0.931	0.944	0.950	0.942	0.952	0.936	0.931
Blue	0.916	0.928	0.931	1.000	0.934	0.908	0.922	0.931	0.929	0.908
Reserves	0.925	0.922	0.944	0.934	1.000	0.930	0.924	0.939	0.935	0.912
Coin	0.942	0.936	0.950	0.908	0.930	1.000	0.926	0.934	0.911	0.909
Rubber	0.922	0.911	0.942	0.922	0.924	0.926	1.000	0.927	0.906	0.914
Diamonds	0.941	0.928	0.952	0.931	0.939	0.934	0.927	1.000	0.916	0.911

Table A2. *Cont.*

	Gold	Silver	Medal	Blue	Res.	Coin	Rubber	Diam.	Tin	Timber
Tin	0.915	0.932	0.936	0.929	0.935	0.911	0.906	0.916	1.000	0.893
Timber	0.908	0.899	0.931	0.908	0.912	0.909	0.914	0.911	0.893	1.000

Table A3. E5-small L2 distance matrix.

	Gold	Silver	Medal	Blue	Res.	Coin	Rubber	Diam.	Tin	Timber
Gold	0.000	0.499	0.487	0.652	0.620	0.540	0.634	0.552	0.656	0.685
Silver	0.499	0.000	0.562	0.610	0.636	0.575	0.684	0.614	0.590	0.723
Medal	0.487	0.562	0.000	0.590	0.534	0.505	0.548	0.500	0.572	0.592
Blue	0.652	0.610	0.590	0.000	0.583	0.687	0.637	0.602	0.606	0.689
Reserves	0.620	0.636	0.534	0.583	0.000	0.602	0.631	0.566	0.582	0.677
Coin	0.540	0.575	0.505	0.687	0.602	0.000	0.622	0.589	0.678	0.685
Rubber	0.634	0.684	0.548	0.637	0.631	0.622	0.000	0.621	0.700	0.671
Diamonds	0.552	0.614	0.500	0.602	0.566	0.589	0.621	0.000	0.664	0.685
Tin	0.656	0.590	0.572	0.606	0.582	0.678	0.700	0.664	0.000	0.744
Timber	0.685	0.723	0.592	0.689	0.677	0.685	0.671	0.685	0.744	0.000

Appendix A.2.2. E5-Base Results

E5-base shows extremely high cosine similarities across all terms, with notable patterns (Table A4). All cosine similarities are above 0.96, suggesting potential overgeneralization. Gold–Diamonds shows the strongest relationship (cosine: 0.992) and L2 distances show more differentiation than cosine similarities (Table A5). Finally, Rubber–Timber shows a surprisingly strong relationship (L2: 1.898).

Table A4. E5-base cosine similarity matrix.

	Gold	Silver	Medal	Blue	Res.	Coin	Rubber	Diam.	Tin	Timber
Gold	1.000	0.983	0.987	0.969	0.988	0.987	0.981	0.992	0.983	0.986
Silver	0.983	1.000	0.969	0.986	0.984	0.975	0.991	0.986	0.985	0.990
Medal	0.987	0.969	1.000	0.960	0.979	0.984	0.967	0.981	0.971	0.974
Blue	0.969	0.986	0.960	1.000	0.975	0.960	0.978	0.969	0.965	0.976
Reserves	0.988	0.984	0.979	0.975	1.000	0.979	0.985	0.985	0.978	0.988
Coin	0.987	0.975	0.984	0.960	0.979	1.000	0.976	0.982	0.986	0.980
Rubber	0.981	0.991	0.967	0.978	0.985	0.976	1.000	0.984	0.987	0.995
Diamonds	0.992	0.986	0.981	0.969	0.985	0.982	0.984	1.000	0.984	0.988
Tin	0.983	0.985	0.971	0.965	0.978	0.986	0.987	0.984	1.000	0.988
Timber	0.986	0.990	0.974	0.976	0.988	0.980	0.995	0.988	0.988	1.000

Table A5. E5-base L2 distance matrix.

	Gold	Silver	Medal	Blue	Res.	Coin	Rubber	Diam.	Tin	Timber
Gold	0.000	3.262	2.804	4.327	2.741	2.817	3.559	2.243	3.296	3.020
Silver	3.262	0.000	4.427	3.066	3.227	3.981	2.401	2.997	3.131	2.527
Medal	2.804	4.427	0.000	4.873	3.585	3.120	4.650	3.431	4.324	4.090
Blue	4.327	3.066	4.873	0.000	3.986	4.893	3.909	4.422	4.699	3.961
Reserves	2.741	3.227	3.585	3.986	0.000	3.656	3.123	3.083	3.759	2.779
Coin	2.817	3.981	3.120	4.893	3.656	0.000	3.926	3.328	2.954	3.530
Rubber	3.559	2.401	4.650	3.909	3.123	3.926	0.000	3.237	2.872	1.898
Diamonds	2.243	2.997	3.431	4.422	3.083	3.328	3.237	0.000	3.196	2.738
Tin	3.296	3.131	4.324	4.699	3.759	2.954	2.872	3.196	0.000	2.770
Timber	3.020	2.527	4.090	3.961	2.779	3.530	1.898	2.738	2.770	0.000

Appendix A.2.3. MiniLM-L6 Results

MiniLM-L6 shows more differentiated relationships than the E5 models (see cosine similarity and L2 distance in Table A6 and Table A7, respectively). Clear clustering of precious metals (Gold–Silver: 0.832) is observed, and lower overall similarities suggest better discrimination. Timber consistently shows the lowest similarities with metals, and L2 distances align well with semantic relationships.

Table A6. MiniLM-L6 cosine similarity matrix.

	Gold	Silver	Medal	Blue	Res.	Coin	Rubber	Diam.	Tin	Timber
Gold	1.000	0.832	0.763	0.685	0.694	0.776	0.609	0.696	0.570	0.542
Silver	0.832	1.000	0.776	0.647	0.563	0.682	0.546	0.660	0.648	0.449
Medal	0.763	0.776	1.000	0.643	0.653	0.688	0.581	0.655	0.572	0.511
Blue	0.685	0.647	0.643	1.000	0.585	0.619	0.569	0.606	0.601	0.424
Reserves	0.694	0.563	0.653	0.585	1.000	0.723	0.577	0.593	0.438	0.470
Coin	0.776	0.682	0.688	0.619	0.723	1.000	0.644	0.664	0.610	0.453
Rubber	0.609	0.546	0.581	0.569	0.577	0.644	1.000	0.615	0.562	0.582
Diamonds	0.696	0.660	0.655	0.606	0.593	0.664	0.615	1.000	0.560	0.468
Tin	0.570	0.648	0.572	0.601	0.438	0.610	0.562	0.560	1.000	0.404
Timber	0.542	0.449	0.511	0.424	0.470	0.453	0.582	0.468	0.404	1.000

Table A7. MiniLM-L6 L2 distance matrix.

	Gold	Silver	Medal	Blue	Res.	Coin	Rubber	Diam.	Tin	Timber
Gold	0.000	0.786	0.849	1.026	1.009	0.838	1.112	1.031	1.217	1.225
Silver	0.786	0.000	0.901	1.157	1.284	1.076	1.287	1.154	1.170	1.436
Medal	0.849	0.901	0.000	1.091	1.072	0.987	1.150	1.097	1.213	1.265
Blue	1.026	1.157	1.091	0.000	1.217	1.139	1.217	1.210	1.213	1.427
Reserves	1.009	1.284	1.072	1.217	0.000	0.969	1.201	1.228	1.435	1.366
Coin	0.838	1.076	0.987	1.139	0.969	0.000	1.073	1.092	1.171	1.353
Rubber	1.112	1.287	1.150	1.217	1.201	1.073	0.000	1.172	1.246	1.188
Diamonds	1.031	1.154	1.097	1.210	1.228	1.092	1.172	0.000	1.296	1.397
Tin	1.217	1.170	1.213	1.213	1.435	1.171	1.246	1.296	0.000	1.473
Timber	1.225	1.436	1.265	1.427	1.366	1.353	1.188	1.397	1.473	0.000

Appendix A.2.4. BGE-Small Results

BGE-small demonstrates balanced semantic understanding. There are strong relationships between related metals (Silver–Coin: 0.794), and moderate cross-category similarities. The model provides a clear distinction between metal and non-metal terms . Also, we find consistent L2 distances supporting cosine similarity patterns (see cosine similarity and L2 distance in Table A8 and Table A9, respectively).

Table A8. BGE-small cosine similarity matrix.

	Gold	Silver	Medal	Blue	Res.	Coin	Rubber	Diam.	Tin	Timber
Gold	1.000	0.783	0.718	0.637	0.621	0.733	0.646	0.750	0.654	0.626
Silver	0.783	1.000	0.711	0.751	0.635	0.794	0.633	0.736	0.742	0.629
Medal	0.718	0.711	1.000	0.619	0.618	0.646	0.566	0.668	0.613	0.606
Blue	0.637	0.751	0.619	1.000	0.554	0.678	0.698	0.696	0.633	0.643
Reserves	0.621	0.635	0.618	0.554	1.000	0.644	0.535	0.568	0.593	0.643
Coin	0.733	0.794	0.646	0.678	0.644	1.000	0.646	0.669	0.731	0.562
Rubber	0.646	0.633	0.566	0.698	0.535	0.646	1.000	0.622	0.605	0.615
Diamonds	0.750	0.736	0.668	0.696	0.568	0.669	0.622	1.000	0.627	0.619

Table A8. *Cont.*

	Gold	Silver	Medal	Blue	Res.	Coin	Rubber	Diam.	Tin	Timber
Tin	0.654	0.742	0.613	0.633	0.593	0.731	0.605	0.627	1.000	0.619
Timber	0.626	0.629	0.606	0.643	0.643	0.562	0.615	0.619	0.619	1.000

Table A9. BGE-small L2 distance matrix.

	Gold	Silver	Medal	Blue	Res.	Coin	Rubber	Diam.	Tin	Timber
Gold	0.000	1.076	1.244	1.425	1.475	1.199	1.363	1.156	1.369	1.449
Silver	1.076	0.000	1.258	1.179	1.446	1.051	1.385	1.187	1.180	1.440
Medal	1.244	1.258	0.000	1.475	1.497	1.397	1.528	1.350	1.466	1.505
Blue	1.425	1.179	1.475	0.000	1.629	1.345	1.288	1.304	1.439	1.444
Reserves	1.475	1.446	1.497	1.629	0.000	1.433	1.618	1.573	1.536	1.461
Coin	1.199	1.051	1.397	1.345	1.433	0.000	1.366	1.334	1.210	1.572
Rubber	1.363	1.385	1.528	1.288	1.618	1.366	0.000	1.408	1.447	1.457
Diamonds	1.156	1.187	1.350	1.304	1.573	1.334	1.408	0.000	1.420	1.462
Tin	1.369	1.180	1.466	1.439	1.536	1.210	1.447	1.420	0.000	1.469
Timber	1.449	1.440	1.505	1.444	1.461	1.572	1.457	1.462	1.469	0.000

Appendix A.2.5. MiniLM-L12-v2 Results

MiniLM-L12-v2 shows the most distinctive differentiation with the strongest distinction between metal and non-metal terms. Gold–Silver maintains the highest similarity (0.732) among all pairs, and there are very low similarities for unrelated pairs (Timber–Coin: 0.144) (Table A10). L2 distances show the largest range, indicating strong discrimination capability (Table A11).

Summarizing all these comments, we can see that each of these advanced models provides different features, and none of them could be assumed to be "better" compared to the others. Our set-word embedding provides another point of view, as is the case with Word2Vec, but the main advantage of our purpose is that we can finally give an interpretation of why the numerical marks are observed.

Table A10. MiniLM-L12-v2 cosine similarity matrix.

	Gold	Silver	Medal	Blue	Res.	Coin	Rubber	Diam.	Tin	Timber
Gold	1.000	0.732	0.480	0.439	0.395	0.645	0.208	0.643	0.422	0.292
Silver	0.732	1.000	0.451	0.418	0.397	0.539	0.264	0.457	0.491	0.297
Medal	0.480	0.451	1.000	0.278	0.350	0.409	0.187	0.353	0.432	0.226
Blue	0.439	0.418	0.278	1.000	0.249	0.371	0.300	0.276	0.433	0.394
Reserves	0.395	0.397	0.350	0.249	1.000	0.401	0.254	0.341	0.433	0.194
Coin	0.645	0.539	0.409	0.371	0.401	1.000	0.335	0.376	0.471	0.144
Rubber	0.208	0.264	0.187	0.300	0.254	0.335	1.000	0.291	0.516	0.406
Diamonds	0.643	0.457	0.353	0.276	0.341	0.376	0.291	1.000	0.439	0.268
Tin	0.422	0.491	0.432	0.433	0.433	0.471	0.516	0.439	1.000	0.438
Timber	0.292	0.297	0.226	0.394	0.194	0.144	0.406	0.268	0.438	1.000

Table A11. MiniLM-L12-v2 L2 distance matrix.

	Gold	Silver	Medal	Blue	Res.	Coin	Rubber	Diam.	Tin	Timber
Gold	0.000	1.693	2.328	2.366	2.641	2.047	3.303	2.399	2.208	2.879
Silver	1.693	0.000	2.353	2.366	2.601	2.303	3.152	2.884	2.036	2.829
Medal	2.328	2.353	0.000	2.593	2.668	2.574	3.275	3.107	2.092	2.930
Blue	2.366	2.366	2.593	0.000	2.805	2.602	3.001	3.226	2.009	2.545

Table A11. *Cont.*

	Gold	Silver	Medal	Blue	Res.	Coin	Rubber	Diam.	Tin	Timber
Reserves	2.641	2.601	2.668	2.805	0.000	2.711	3.261	3.227	2.271	3.134
Coin	2.047	2.303	2.574	2.602	2.711	0.000	3.108	3.165	2.242	3.261
Rubber	3.303	3.152	3.275	3.001	3.261	3.108	0.000	3.554	2.455	2.933
Diamonds	2.399	2.884	3.107	3.226	3.227	3.165	3.554	0.000	2.790	3.412
Tin	2.208	2.036	2.092	2.009	2.271	2.242	2.455	2.790	0.000	2.288
Timber	2.879	2.829	2.930	2.545	3.134	3.261	2.933	3.412	2.288	0.000

References

1. Clark, S. Vector space models of lexical meaning. In *The Handbook of Contemporary Semantics*; Lappin, S., Fox, C., Eds.; Blackwell: Malden, MA, USA, 2015; pp. 493–522. [CrossRef]
2. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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, Minneapolis, MN, USA, 2–7 June 2019; Burstein, J., Doran, C., Solorio, T., Eds.; Volume 1, pp. 4171–4186. [CrossRef]
3. Incitti, F.; Urli, F.; Snidaro, L. Beyond word embeddings: A survey. *Inf. Fusion* **2023**, *89*, 418–436. [CrossRef]
4. Mikolov, T.; Chen, K.; Corrado, G.S.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
5. Pennington, J.; Socher, R.; Manning, C. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543. [CrossRef]
6. Radford, A.; Narasimhan, K. Improving Language Understanding by Generative Pre-Training. Preprint. 2018. Available online: https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035 (accessed on 23 December 2024).
7. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762. [CrossRef]
8. Grand, G.; Blank, I.A.; Pereira, F.; Fedorenko, E. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nat. Hum. Behav.* **2022**, *6*, 975–987. [CrossRef]
9. Manetti, A.; Ferrer-Sapena, A.; Sánchez-Pérez, E.A.; Lara-Navarra, P. Design Trend Forecasting by Combining Conceptual Analysis and Semantic Projections: New Tools for Open Innovation. *J. Open Innov. Technol. Mark. Complex.* **2021**, *7*, 92. [CrossRef]
10. Zadeh, L.A. Quantitative fuzzy semantics. *Inf. Sci.* **1971**, *3*, 159–176. [CrossRef]
11. Zadeh, L.A. A Fuzzy-Set-Theoretic Interpretation of Linguistic Hedges. *J. Cybern.* **1972**, *2*, 4–34. [CrossRef]
12. Saranya, M.; Amutha, B. A Survey of Machine Learning Technique for Topic Modeling and Word Embedding. In Proceedings of the 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 14–15 March 2024; Volume 1, pp. 1–6.
13. Hongliu, C.A.O. Recent advances in text embedding: A Comprehensive Review of Top-Performing Methods on the MTEB Benchmark. *arXiv* **2024**, arXiv:2406.01607. [CrossRef]
14. Wang, Y.; Mishra, S.; Alipoormolabashi, P.; Kordi, Y.; Mirzaei, A.; Arunkumar, A.; Khashabi, D. Super-naturalinstructions: Generalization via declarative instructions on 1600+ NLP tasks. *arXiv* **2022**, arXiv:2204.07705.
15. Georgila, K. Comparing Pre-Trained Embeddings and Domain-Independent Features for Regression-Based Evaluation of Task-Oriented Dialogue Systems. In Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Kyoto, Japan, 18–20 September 2024; pp. 610–623.
16. Baroni, M.; Zamparelli, R. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, MA, USA, 9–11 October 2010; pp. 1183–1193.
17. Erk, K. Vector space models of word meaning and phrase meaning: A survey. *Lang. Linguist. Compass* **2012**, *6*, 635–653. [CrossRef]
18. Arens, R.F.; Eels, J.J. On embedding uniform and topological spaces. *Pac. J. Math* **1956**, *6*, 397–403. [CrossRef]
19. Kosub, S. A note on the triangle inequality for the Jaccard distance. *Pattern Recognit. Lett.* **2019**, *120*, 36–38. [CrossRef]
20. Deza, M.M.; Deza, E. *Encyclopedia of Distances*, 1st ed.; Springer: Berlin/Heidelberg, Germany, 2009. [CrossRef]
21. Gardner, A.; Kanno, J.; Duncan, C.A.; Selmic, R. Measuring distance between unordered sets of different sizes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 137–143. [CrossRef]
22. Cobzaş, C. *Functional Analysis in Asymmetric Normed Spaces*; Springer Science & Business Media: Berlin, Germany, 2012. [CrossRef]

23. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–8 December 2013; pp. 3111–3119. [CrossRef]
24. Mikolov, T.; Yih, W.T.; Zweig, G. Linguistic regularities in continuous space word representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, GA, USA, 9–14 June 2013; pp. 746–751.
25. Candido, L.; Cúth, M.; Doucha, M. Isomorphisms between spaces of Lipschitz functions. *J. Funct. Anal.* **2019**, *277*, 2697–2727. [CrossRef]
26. Cobzaş, C.; Miculescu, R.; Nicolae, A. *Lipschitz Functions*; Springer: Berlin, Germany, 2019. [CrossRef]
27. Erdogan, E.; Ferrer-Sapena, A.; Jimenez-Fernandez, E.; Sánchez Pérez, E. Index spaces and standard indices in metric modelling. *Nonlinear Anal. Model. Control* **2022**, *27*, 1–20. [CrossRef]
28. Kuratowski, C. Quelques problèmes concernant les espaces métriques non-séparables. *Fundam. Math.* **1935**, *25*, 534–545. [CrossRef]
29. Ruas, T.; Grosky, W. Keyword extraction through contextual semantic analysis of documents. In Proceedings of the 9th International Conference on Management of Digital EcoSystems, Bangkok, Thailand, 7–10 November 2017; pp. 150–156. [CrossRef]
30. Shi, F.; Qing, P.; Yang, D.; Wang, N.; Lei, Y.; Lu, H.; Lin, X.; Li, D. Prompt space optimizing few-shot reasoning success with large language models. *arXiv* **2023**, arXiv:2306.03799.
31. Wan, X.; Sun, R.; Dai, H.; Arik, S.O.; Pfister, T. Better zero-shot reasoning with self-adaptive prompting. *arXiv* **2023**, arXiv:2305.14106.
32. Zheng, C.T.; Liu, C.; Wong, H.S. Corpus-based topic diffusion for short text clustering. *Neurocomputing* **2018**, *275*, 2444–2458. [CrossRef]
33. Song, Y.; Wang, T.; Cai, P.; Mondal, S.K.; Sahoo, J.P. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *Acm Comput. Surv.* **2023**, *55*, 1–40. [CrossRef]
34. Xu, S.; Pang, L.; Shen, H.; Cheng, X.; Chua, T.S. Search-in-the-Chain: Interactively Enhancing Large Language Models with Search for Knowledge-intensive Tasks. In Proceedings of the ACM on Web Conference 2024, Singapore, 13–17 May 2024; pp. 1362–1373.
35. Xiong, H.; Bian, J.; Li, Y.; Li, X.; Du, M.; Wang, S.; Helal, S. When search engine services meet large language models: Visions and challenges. *IEEE Trans. Serv. Comput.* **2024**, *17*. [CrossRef]
36. Bodria, F.; Giannotti, F.; Guidotti, R.; Naretto, F.; Pedreschi, D.; Rinzivillo, S. Benchmarking and Survey of Explanation Methods for Black Box Models. *Data Min. Knowl. Disc.* **2023**, 37, 1719–1778. [CrossRef]
37. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
38. Lyu, Q.; Apidianaki, M.; Callison-Burch, C. Towards Faithful Model Explanation in NLP: A Survey. *Comput. Linguist.* **2024**, *50*, 657–723. [CrossRef]
39. Wang, L.; Yang, N.; Huang, X.; Yang, L.; Majumder, R.; Wei, F. Multilingual E5 Text Embeddings: A Technical Report. *arXiv* **2024**, arXiv:2402.05672.
40. Wang, W.; Wei, F.; Dong, L.; Bao, H.; Yang, N.; Zhou, M. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020; Volume 33, pp. 5776–5788.
41. Lin, Y.; Ding, B.; Jagadish, H.V.; Zhou, J. SMARTFEAT: Efficient Feature Construction through Feature-Level Foundation Model Interactions. *arXiv* **2023**, arXiv:2309.07856.