applied sciences

MDPI

Article

WERECE: An Unsupervised Method for Educational Concept Extraction Based on Word Embedding Refinement

Jingxiu Huang, Ruofei Ding, Xiaomin Wu, Shumin Chen, Jiale Zhang, Lixiang Liu and Yunxiang Zheng *

School of Educational Information Technology, South China Normal University, No. 55 Western Zhongshan Avenue, Guangzhou 510631, China; jimsow@m.scnu.edu.cn (J.H.); ruofei@m.scnu.edu.cn (R.D.); 201928210238@m.scnu.edu.cn (X.W.); 20192831014@m.scnu.edu.cn (S.C.); 20222831048@m.scnu.edu.cn (J.Z.); 20222821033@m.scnu.edu.cn (L.L.)

* Correspondence: dr.zheng.scnu@hotmail.com

Abstract: The era of educational big data has sparked growing interest in extracting and organizing educational concepts from massive amounts of information. Outcomes are of the utmost importance for artificial intelligence–empowered teaching and learning. Unsupervised educational concept extraction methods based on pre-trained models continue to proliferate due to ongoing advances in semantic representation. However, it remains challenging to directly apply pre-trained large language models to extract educational concepts; pre-trained models are built on extensive corpora and do not necessarily cover all subject-specific concepts. To address this gap, we propose a novel unsupervised method for educational concept extraction based on word embedding refinement (i.e., word embedding refinement–based educational concept extraction (WERECE)). It integrates a manifold learning algorithm to adapt a pre-trained model for extracting educational concepts while accounting for the geometric information in semantic computation. We further devise a discriminant function based on semantic clustering and Box–Cox transformation to enhance WERECE's accuracy and reliability. We evaluate its performance on two newly constructed datasets, EDU-DT and EDUTECH-DT. Experimental results show that WERECE achieves an average precision up to 85.9%, recall up to 87.0%, and F1 scores up to 86.4%, which significantly outperforms baselines (TextRank, term frequency–inverse document frequency, isolation forest, K-means, and one-class support vector machine) on educational concept extraction. Notably, when WERECE is implemented with different parameter settings, its precision and recall sensitivity remain robust. WERECE also holds broad application prospects as a foundational technology, such as for building discipline-oriented knowledge graphs, enhancing learning assessment and feedback, predicting learning interests, and recommending learning resources.

Keywords: concept extraction; word embedding; semantic computation; knowledge graph; manifold learning; clustering

Citation: Huang, J.; Ding, R.; Wu, X.; Chen, S.; Zhang, J.; Liu, L.; Zheng, Y. WERECE: An Unsupervised Method for Educational Concept Extraction Based on Word Embedding Refinement. *Appl. Sci.* **2023**, *13*, 12307. https://doi.org/10.3390/app132212307

Academic Editor: Andrea Prati

Received: 22 October 2023
Revised: 11 November 2023
Accepted: 13 November 2023
Published: 14 November 2023

1. Introduction

The expansion of the internet and new learning media, coupled with an information explosion, has brought seemingly limitless knowledge enrichment [1]. The era of big data has rendered it necessary to extract and organize domain knowledge from a vast amount of information; doing so lessens people's likelihood of becoming disoriented when working online. Experts initially extracted conceptual knowledge manually from domain texts. This process promoted information comprehension and dissemination [2]. However, manual extraction requires substantial time and effort, and different experts may possess varied understandings of the same concept. The growth of the internet has meant conventional concept extraction no longer meets the demand for handling large volumes of online data or updating domain knowledge in a timely fashion. Considerable research has therefore aimed to develop more streamlined, sophisticated techniques for automated concept extraction [3,4].

The generic paradigms of supervised and unsupervised approaches for automated concept extraction are valuable [2,4,5]. A sufficient number of pre-annotated training samples are necessary for supervised extraction. This method converts the concept extraction problem into a classification problem, where a classification model is established to categorize numerous concepts. Yet this technique calls for annotating data by hand, with the quality of annotation directly affecting models' extraction performance. Another paradigm, unsupervised concept extraction, consists of rule-based methods [6,7], dictionary-based methods [8], statistical methods [9,10], and semantic-based methods [11,12]. Each category has benefits and drawbacks. For instance, dictionary-based methods are known for their speed and effectiveness, owing to well-constructed dictionaries. However, they tend to disregard important concepts and may produce erroneous outcomes. Statistical methods are comparatively efficient, cost-effective, and objectively reliable. Nonetheless, these approaches cannot accurately extract all target domain concepts due to inherent fuzziness and ambiguity in their intrinsic meaning and boundaries; such problems are even more pronounced for noisy datasets. To mitigate these issues in domain concept extraction, semantic information has been used to normalize concepts (i.e., by aligning extracted concepts with a domain-specific context). This strategy enhances the accuracy, coverage, and interoperability of domain concept extraction [11–13]. Briefly, automated concept extraction is a multifaceted and arduous task in which several techniques are combined to refine task performance. Few studies have sought to improve concept extraction methods' explainability and robustness. Further exploration is thus needed to improve learning and instruction in educational settings (e.g., online learning and intelligent education).

In the education field, foundational subject-related knowledge reflects concepts frequently expressed as single words or phrases. Educational concepts must be blended into learning and instructional practices; students' inadequate conceptual understanding can cause them to forget information during the learning process [14,15]. Similarly, instructors can enhance learning materials' quality through a clear sense of subject-specific concepts [16,17]. Extracting concepts from various subject areas based on extensive unstructured text is thus critical for instructors and students, especially to enrich teaching and learning activities. Yet educational concept extraction faces certain challenges. Firstly, the educational domain consists of a range of topics and bodies of knowledge, each featuring distinct concepts and terminology. Accurately extracting educational concepts requires an in-depth understanding of the subject matter as well as contextual awareness. Secondly, these concepts are often polysemous and nuanced, such that one concept can have multiple meanings and usages. Concepts also tend to appear in multiple subject areas (e.g., in deep learning and information technology). Rich linguistic understanding and semantic computation [18,19] are needed to extract educational concepts effectively. Thirdly, an obstacle within this type of concept extraction entails the scarcity of extensive annotated data. Acquiring large-scale, high-quality annotated data is also labor-intensive. To allay problems in contextual awareness and semantic computation, scholars must examine how to fully represent educational concepts' semantic information. The unsupervised concept extraction methods that researchers typically use mostly rely on word frequency statistics, which can easily overlook low-frequency educational concepts and fail to capture underlying semantic information [16,20,21]. Fortunately, due to the increasing prevalence of pre-trained models, numerous scholars have started to use word embedding techniques for unsupervised concept extraction [18,19,22]. Even so, it is not desirable to directly incorporate pre-trained models into unsupervised concept extraction tasks: these models are trained on extensive corpora and may not cover every subject-specific concept. Therefore, how to refine pre-trained models to pinpoint more semantic information for educational concept extraction is another impressive research gap and merits attention.

To bridge some of these gaps, we present an unsupervised method for extracting educational concepts from unstructured text. Our work makes the following contributions:

(1) A novel unsupervised method is proposed for educational concept extraction based on word embedding refinement. The proposed method, named word embedding

refinement–based educational concept extraction (WERECE), efficiently integrates the semantic information of domain concepts. Its performance surpasses popular baselines and state-of-the-art algorithms in educational concept extraction.

(2) We introduce a manifold learning algorithm to adapt pre-trained large language models to a downstream natural language processing (NLP) task. The algorithm fully considers geometric information in semantic computation and reinforces semantic clustering among educational concepts.

(3) A discriminant function based on semantic clustering and Box–Cox transformation is developed to improve the accuracy and reliability of educational concept extraction.

(4) Two real-world datasets are created for educational concept extraction and used to experimentally assess WERECE's effectiveness. We also evaluate how WERECE's parameter settings influence its performance.

The rest of this paper is organized as follows. Section 2 gives an overview of studies on concept extraction and its educational applications. Section 3 describes WERECE and the method's main steps. Section 4 discusses our experiments and analysis of WERECE's performance in terms of precision, recall, and F scores. Finally, Section 5 presents conclusions and suggested future work.

2. Related Work

Various studies have examined concept extraction based on extensive sets of textual material. The popularity of online learning and new learning media has ignited concerns about concept extraction in educational settings. As such, Section 2.1 outlines ways to extract concepts from text without a predetermined application. Section 2.2 reviews concept extraction methods that target the educational domain.

2.1. Generic Concept Extraction Methods

Concept extraction, a core NLP task, refers to identifying and extracting predefined concepts or patterns from textual data. This task is difficult given the complex and dynamic nature of language. Scholars have investigated concept extraction in numerous areas, such as clinical medicine, information retrieval, and automation engineering. A recent review of the clinical literature on this topic indicated that most approaches to extracting domain concepts from clinical text fall into two categories: (1) rule-based methods and (2) machine learning methods, including deep learning approaches and hybrid approaches [4]. Likewise, Kang et al. suggested that concept extraction strategies related to information retrieval can be classified as either machine learning methods, corpus-based methods, glossary-based methods, or heuristic-based methods [23]. Currently, concept extraction often involves either supervised or unsupervised strategies. These approaches normally follow a four-step procedure of preprocessing, generating a list of candidate concepts, identifying concepts from candidates, and evaluating those concepts.

Our work is an example of unsupervised concept extraction, a task that can be further divided into four groups (summarized in Table 1): rule-based methods, dictionary-based methods, statistical methods, and semantic-based methods. First, rules and patterns in rule-based methods are predefined to extract concepts from text. Regular expressions or pattern-matching techniques are prevalent. Rule-based concept extraction adheres to grammatical rules, semantic rules, and related aspects to process a corpus and extract multi-character units that conform to predefined rules. These units are eventually labeled as concepts. Szwed employed a rule-based method that involved transforming detected names according to Polish grammar rules, utilizing a user-friendly approach for specifying transformation patterns through annotations to extract concepts from unstructured Polish texts [6]. Stanković et al. developed a rule-based approach that relies on a system of language resources to tackle the multi-word term problem in domain concept extraction [7]. One benefit of rule-based methods is their capacity to manage patterns and implement domain-specific knowledge. Nevertheless, these approaches are labor-intensive and time-consuming to develop. They also may not capture linguistic diversity. Second,

dictionary-based concept extraction methods use pre-defined dictionaries of concept keywords to extract relevant information. Words or phrases in the target text are compared with dictionary entries via similarity metrics or string-matching algorithms [8,24]. These techniques are faster than rule-based methods but heavily depend on the dictionary's quality and coverage; they may include noise or miss concepts that are absent from the dictionary. Third, statistical methods emphasize modeling and analyzing potential patterns among domain concepts in a target text. Statistical metrics such as term frequency–inverse document frequency (TF–IDF), co-occurrence, and neighbors are popular heuristics when ranking candidate concepts. The TF–IDF method is premised on the fact that domain-specific concepts exhibit much higher frequencies in some domains than in others, akin to the word frequency patterns provided by TF–IDF [9]. Candidate domain concepts can also be ranked statistically by depicting the extracted concepts as nodes on a graph and appraising their roles using network properties such as concept centrality and connectivity. Concepts that occupy more prominent positions within the graph receive higher scores, reflecting the representativeness of both the node and the concept [5]. Graph-based methods for concept extraction include the TextRank [10] approach and its variations, such as Ne-rank [25], TopicRank [26], and MultipartiteRank [27]. Even though TF–IDF and TextRank each disregard contextual information, they were taken as baselines in our experiments because they have performed consistently well in past research. Moreover, with advances in deep learning and the emergence of distributed representation models in recent years, scholars have increasingly sought to incorporate word embedding models into concept extraction. Tulkens et al. built an unsupervised clinical concept extraction system using a skip-gram-based embedding model to create concept representations [12]. Xiong et al. used the Bidirectional Encoder Representation from Transformers (BERT) model to improve the performance of the TextRank method [13]. These efforts highlight the importance of concepts' semantic features and their semantic associations. Motivated by such work, we similarly employed word embeddings to acquire semantic representations for candidate concepts and classified them using a novel discriminant function derived from the K-means algorithm.

2.2. Concept Extraction in Education

Concept extraction is a fundamental technique for knowledge mining in education (e.g., when identifying topics in students' online discussions or arranging educational knowledge graphs). Studies have demonstrated that automated domain concept extraction brings deep insights for teaching and learning [14,15,28,29]. Chen et al. identified e-learning domain concepts from academic articles to assemble a concept map; this helped teachers create adaptive learning materials and enabled students to better grasp the complete picture of subject knowledge [16]. Conde formulated a tool to ascertain terms from electronic textbooks and assist teachers in crafting instructional materials [17]. Peng et al. extracted topic concepts from students' forum posts, enabling instructors to detect and trace students' learning engagement with discourse content [30]. A set of concepts extracted from subject materials, along with a group of association rules, can be used to construct knowledge graphs and thereby promote teaching and learning. A systematic review revealed that the relationships among domain concepts are essential for estimating or predicting learners' knowledge states [15]. Together, such research has shown concept extraction to be crucial in teaching and learning practices. However, popular approaches in educational studies (e.g., TF–IDF and latent Dirichlet allocation) depend on word frequency statistics, which can easily overlook low-frequency educational concepts and struggle to capture the semantic information behind text. Therefore, it is imperative to determine how to exploit semantic information from educational concepts to facilitate concept extraction.

Many strategies adopted in educational settings involve TF–IDF, C/NC values, and graph-based ranking. These statistical approaches to concept extraction (i.e., from textual data) are generally contingent on word frequency or key words. Lin Zhang proposed a hybrid method based on the TextRank algorithm and TF–IDF for key concept extraction

and sentiment analysis of educational texts [21]. Liu improved the Chinese term extraction method by using C/NC values [20]. Although statistical methods are applicable to concept extraction, they traditionally require extensive domain knowledge and labeling to identify meaningful features. In contrast, word embedding techniques can learn directly from text corpora without manual labeling or feature engineering; that is, they can learn in an unsupervised manner. Each dimension of word embeddings can also reflect certain aspects of lexical meaning, thereby providing rich semantic information [31].

Table 1. A summary of generic concept extraction methods.

Methods	Core Processes	Strengths	Weaknesses	Articles
Rule-based methods	▪ Adhering to grammatical rules, semantic rules, and related aspects to process a corpus; ▪ Extracting multi-character units that conform to predefined rules and are labeled as concepts.	▪ Having the ability to manage well-defined patterns; ▪ Supporting domain-specific knowledge. ▪ Enabling access and providing transparency of the concept extraction process.	▪ Labor-intensive; ▪ Time-consuming; ▪ Disregards linguistic diversity.	[6,7]
Dictionary-based methods	▪ Using a predefined concept dictionary to compare words or phrases in the text with dictionary entries; ▪ Employing similarity metrics or string-matching algorithms to extract relevant information.	▪ Allowing faster implementation; ▪ Allowing for quick extensions for domain adaption; ▪ Providing scalability to large datasets.	▪ Depends on dictionary quality and coverage; ▪ Missing out-of-dictionary concepts; ▪ Underperforming in concept extraction.	[8,24]
Statistical-based methods	▪ Analyzing word frequency and co-occurrence patterns; ▪ Ranking based on statistical features to identify concepts in the text (e.g., weight-based ranking, graph-based ranking).	▪ Having ability to model potential patterns among domain concepts; ▪ Allowing highly customized extracting domain concepts; ▪ Being robust to noise.	▪ Being sensitive to data quality; ▪ Disregarding contextual information; ▪ Missing semantic associations.	[9,25–27]
Semantic-based methods	▪ Using predefined grammar rules to identify candidate concepts; ▪ Utilizing pretrained models to obtain semantic vectors for candidates; ▪ Applying post-processing techniques to determine target concepts from the text.	▪ Obtaining higher precision and recall scores by capturing deeper meaning and contextual information in concept extraction; ▪ Taking advantage of the state-of-the-art NLP techniques (e.g., word embeddings, BERT); ▪ Having scalability to different domains.	▪ Relying on the quality and coverage of word embeddings; ▪ Requiring larger computational resources; ▪ Having challenges in evaluation and explainability.	[12,13]

At present, pre-trained large language models can obtain word representations with more semantic information and have been employed for educational concept extraction. Pan et al. extended the pre-trained embedding model by adding a graph propagation algorithm to capture relationships between words and courses, enabling domain concepts

to be identified within a course [18]. Albahr et al. used the skip-gram model with the Wikipedia corpus to ascertain word embedding vectors for concept extraction in massive open online courses [19]. To address noisy and incomplete annotations during high-quality knowledgeable concept extraction from these types of courses, Lu et al. developed a three-stage framework [22]. It harnessed pre-trained language models explicitly and implicitly and integrated discipline-embedding models with a self-training strategy. These models are usually trained on large-scale corpora, making them highly robust and able to implicitly encode real knowledge concepts [32]. However, when using pre-trained models for concept extraction, the generality of corpora may cause extracted concepts not to match the semantics in a specific domain. Put simply, this method's feasibility is limited in domain-specific concept extraction in the absence of extensive and high-quality domain-specific corpora. Publicly available pre-trained word embedding models are sufficient for NLP tasks. Researchers from different fields have since fine-tuned these models on target domain texts to improve performance in downstream NLP tasks. In the legal domain, Chalkidis et al. developed the Legal-BERT model based on BERT and realized higher performance [33]. Wang et al. showed that word embeddings trained on biomedical corpora captured the semantics of medical terms better than word embeddings trained on general domain corpora [34]. Clavi and Gal noted that domain-specific large pre-trained models could have promising results for learning analytics [35]. Concept extraction performance can thus be enhanced by optimizing pre-trained models. The true test lies in effectively incorporating pre-trained models tailored to domain-specific semantics into educational concept extraction.

3. Methods

An unsupervised method for educational concept extraction, named WERECE, was proposed to maximize the use of semantic information in pre-trained word embedding models. The two-phase procedure is illustrated in Figure 1: model training (solid lines) is followed by model prediction (dotted lines). In the training phase, a collection of seed concepts chosen from the target domain was projected onto word embeddings for word representation. The generated word embedding vectors of seed concepts were subsequently refined through a manifold learning model. The refined vectors served as input for the K-means clustering algorithm to determine domain concepts' cluster centers. Building on these cluster centers, we next developed a distance-based discriminant function. In the prediction phase, state-of-the-art NLP techniques were adopted in a loose candidate generation step. Then candidate concepts were transformed into word embedding vectors followed by word re-embedding with the manifold learning model. Finally, candidate concepts' refined vectors were fed into the discriminant function to uncover domain concepts.

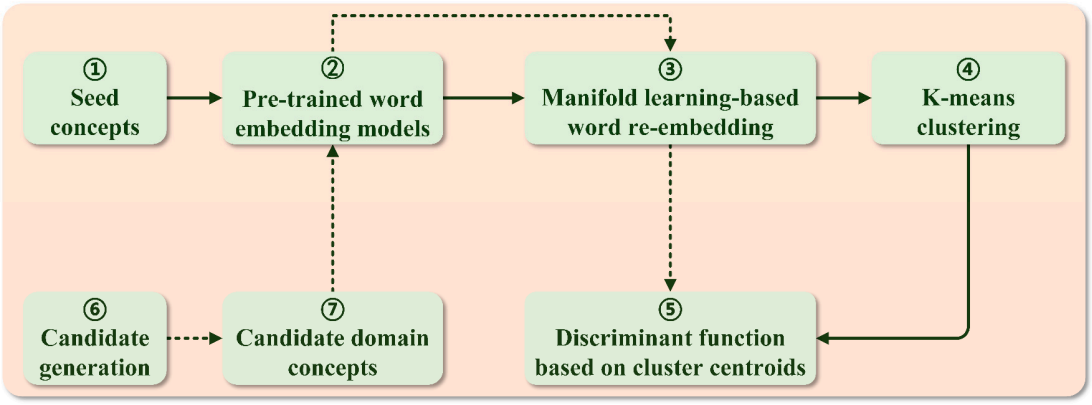

Figure 1. Workflow of our WERECE method.

3.1. Domain Concept Representation with Pre-Trained Word Embeddings

Semantic information is incredibly effective in domain concept extraction. For instance, to represent semantic information within words and phrases, word embeddings are now preferred for various NLP applications [36,37]. Embeddings like word2vec [38] adhere to the principle that words with similar meanings are likely to appear in similar contexts. Based on this idea, word embeddings have been pre-trained on huge volumes of natural language text. This step enables researchers to incorporate more abundant contextual information into representation vectors and thus increase downstream NLP task efficiency.

The distributed representation of domain concepts was initialized using a pre-trained word embedding model released by Tencent AI Lab [39]. The selected model provides a 200-dimension vector representation for a large number of domain words and phrases in Chinese. Its superiority in different NLP tasks is attributable to its large-scale data and to a well-designed training algorithm that accounts for word co-occurrence patterns and their relative positions. Despite the merit of Tencent embedding, two obstacles remained to be addressed when we used it to produce representation vectors for domain concepts. First, a few domain concepts—especially in phrases—resulted in the out-of-vocabulary (OOV) problem that can accompany pre-trained word embeddings. Second, loading the entire Tencent embedding program routinely requires substantial memory resources and time. We therefore used Magnitude [40], a utility package in Python, to quickly process the Tencent embedding. Magnitude handles the OOV issue for any embedding model, particularly for those without internal OOV support.

3.2. Manifold Learning-Based Word Re-Embedding for Domain Concepts

Most distributed word representation models ignore how words' geometric structure affects semantic calculation [41–43]. Word re-embedding seeks to eliminate this oversight by refining the word representation based on intrinsic geometric information in the original embedding space. Scholars have deployed manifold learning algorithms for this purpose by integrating geometric information between words and their neighbors. To get the most out of semantic information for domain-specific concept extraction, both local and global geometric information were exploited to refine seed concept vectors originating from pre-trained word embeddings through a manifold learning algorithm.

IsoMap, short for "isometric mapping", is a popular approach to manifold learning [44]. This method extends multidimensional scaling by introducing geodesic distance to measure the similarity between all data points. Compared to the locally linear embedding algorithm that has been used for word embedding refinement, IsoMap captures global information; it uses local information to construct a global neighborhood graph that describes the original embedding space's overall structure. This algorithm can hence be characterized by a neighborhood graph and geodesic distance. Given N seed concept vectors $V = \{v_1, v_2, v_3, \cdots, v_N\}$, the IsoMap algorithm was implemented in four steps. To start, K nearest neighbor methods (e.g., Ball-tree, KD-tree) were applied to search for neighbors for each concept vector. Next, an adjacency matrix $A_{N\times N}$ was created to describe undirected neighborhood relationships among vectors. Each element a_{ij} in $A_{N\times N}$ denotes the connection weight between concept vectors v_i and v_j. If a neighborhood relationship exists between v_i and v_j, then a_{ij} is initially determined by the Euclidean metric; otherwise, a_{ij} is set to infinity. Thereafter, each matrix element a_{ij} was updated to approximate the geodesic distance from v_i to v_j. This approximation was based on the Floyd algorithm, a shortest-path algorithm grounded in dynamic programing as shown in Equation (1). Lastly, the adjacency matrix $A_{N\times N}$ with final distance values was plugged into multidimensional scaling, and eigendecomposition was performed to construct a refined embedding space U with d dimensions. This refined space ensured that the intrinsic geometry of the original spatial data was best preserved in low dimensions. Suppose that the seed concept vectors projected onto the refined embedding space can be represented as $R = \{r_1, r_2, r_3, \cdots, r_N\}$. Then, multidimensional scaling can yield an optimal projection by minimizing the cost function in Equation (2). The solution to the global minimization problem is identical to

Equation (3), where Σ denotes the top d eigenvalues of $A^T A$, and V is the eigenvector corresponding to $A^T A$.

$$a_{ij} = \underbrace{\min\left(a_{ij},\ a_{ik} + a_{kj}\right)}_{i,\ j,\ k \in \{1,2,3,\cdots,\ N\}}. \tag{1}$$

$$J(R) = \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\|r_i - r_j\|_2^2 - a_{ij} \right)^2. \tag{2}$$

$$U = \sqrt{\Sigma} V^T. \tag{3}$$

3.3. K-Means Clustering Algorithm

K-means clustering is an unsupervised learning method that follows a data partition strategy based on Euclidean distance [45]. It is distinguished by an iterative learning process where the center of each cluster (i.e., the centroid, calculated as the mean of data points in the cluster) is continually updated until it meets a convergence criterion. The criterion of inertia abides by the within-cluster sum of squares. K-means aims to minimize a cost function, which is the sum of the squared error on K clusters. This function is mathematically written as Equation (4), where c_k is the k-th cluster, x_i is the i-th data point in cluster c_k, and μ_{c_k} is the centroid of cluster c_k.

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in c_k} \|x_i - \mu_{c_k}\|_2^2. \tag{4}$$

The K-means clustering algorithm has several advantages [45]: (1) relatively high computing efficiency, (2) low time complexity, and (3) good interpretability. It has thus been applied in contexts such as computer vision and image processing, information retrieval, and knowledge extraction. However, the K-means algorithm must randomly select K data points to initialize cluster centroids and can potentially converge to undesired local minima, which increases the randomness and unsteadiness of cluster membership. Therefore, initially chosen cluster centroids and the number of clusters (i.e., K) warrant careful consideration before implementing this algorithm. We built a Python program with the scikit-learn package and selected initial centroids by measuring their probability of contributing to the overall inertia in each sampling step. No ideal heuristic strategy or mathematical criterion was available to determine the value of K. Therefore, we empirically tested different values and evaluated the cluster results based on an intrinsic metric. For the sake of computational efficiency when dealing with extensive data, the Calinski-Harabasz (CH) score [46] was adopted to estimate clustering performance. The CH score can be calculated as in Equation (5), where n_k is the number of members in cluster c_k and μ is the centroid of the entire dataset with given N seed concept vectors. The higher the value of the CH index, the better the clustering validity; that is, clusters are primely separated from each other and are distinctly preferable.

$$\mathrm{CH} = \frac{\left(\sum_{k=1}^{K} n_k \|\mu_{c_k} - \mu\|_2^2 \right)(N - K)}{\left(\sum_{k=1}^{K} \sum_{i=1}^{n_k} \|x_i - \mu_{c_k}\|_2^2 \right)(K - 1)}. \tag{5}$$

3.4. Discriminant Function Based on Cluster Centroids

Given the cluster centroids that the K-means algorithm produced, the Euclidean distances of all data points to their cluster centroids were obtained. A discriminant function was then defined to discern whether candidate concepts that had been transformed into refined vectors were involved in a specific domain. Generally, the fitted K-means model inherently determines a new data point (i.e., candidate concept) subjected to the target domain. However, because of outliers and noise, cluster closeness affects this determina-

tion [47]. We assumed that data points' Euclidean distances to cluster centroids aligned with the normal distribution. In the k-cluster c_k displayed in Figure 2, the mean distance $\overline{d_k}$ of its members to its centroid μ_{c_k} and the corresponding standard deviation σ_k can be respectively obtained. The empirical rule of normal distribution indicates that the probability of the distance of data point x_i to μ_{c_k} being less than $\left(\overline{d_k} + 2 \times \sigma_k\right)$ is roughly 0.95. Thus, the discriminant function DF was written as shown in Equation (6), where d^k is the distance of a new data point to the centroid μ_{c_k}. This discrimination is similar to the data points falling outside each cluster being considered outliers [48]. However, the normal distribution is difficult to satisfy in reality, such that data samples' skewness may mislead the discriminant function. To address this challenge, Box–Cox [49] transformation was employed to generalize the Euclidean distances of all data points to their centroids. This technique is a generalized form of power transformation and is formally identical to Equation (7), where d_k is the distance of a data point to the centroid μ_{c_k} and λ is a transformation parameter. Box–Cox transformation can reduce the unobserved error to a certain extent, thereby enhancing data normality, symmetry, and variance equality. This form of transformation has hence been used to improve the accuracy and reliability of data modeling in numerous areas [50–52].

$$DF = \bigcup_{k=1}^{K} \left\{ d^k < \left(\overline{d_k} + 2 \times \sigma_k\right) \right\}. \tag{6}$$

$$F(d_k, \lambda) = \begin{cases} \frac{d_k^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \log(d_k) & \text{if } \lambda = 0. \end{cases} \tag{7}$$

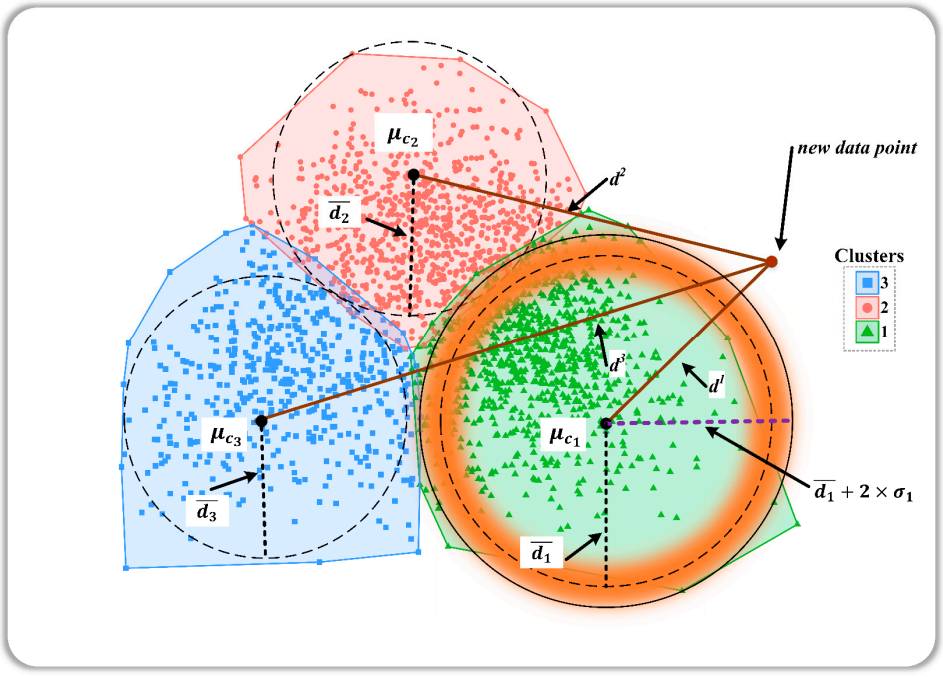

Figure 2. Cluster centroids of educational concepts accounting for the discriminant function. Note that k is the number of clusters that the K-means algorithm produced (cluster 1 is in green, cluster 2 is in red, and cluster 3 is in blue); μ_{c_k} is the centroid of each cluster; $\overline{d_k}$ is the mean distance of each cluster's members to its cluster centroid μ_{c_k}; and d^k is the distance of a new data point to the cluster centroid μ_{c_k}. The probability of the distance of a new data point to the cluster centroid empirically being less than $\left(\overline{d_k} + 2 \times \sigma_k\right)$ is roughly 0.95.

4. Experiments

We addressed the following research questions when assessing the performance of the proposed method:

(1) How feasible is this method for educational concept extraction?
(2) How does this method perform when the dimensions of refined embedding vectors, the number of seed concepts, and the number of clusters change?
(3) How effective is this method for educational concept extraction when compared with baselines?

A series of experiments on two educational datasets was carried out to answer these questions. In the ensuing subsections, we first describe our dataset preparation along with evaluation metrics and baselines. Next, we present the experimental results compared with baseline methods.

4.1. Dataset Preparation

To the best of our knowledge, no gold-standard datasets are publicly available with which to measure domain concept extraction in education. Thus, a pair of education-related datasets was created to proceed with our experiments based on public data and the literature. The first dataset (EDU-DT) was derived from MOOCCube [53], a large-scale data warehouse covering various domains that has been used for different NLP tasks [54]. When assembling this dataset, 2956 educational concepts deemed as positives were screened out from MOOCCube; 1000 concepts from other domains (e.g., world history, computer science, and management science) were randomly chosen as negatives. The EDU-DT dataset was further separated into two parts: (1) a training set with 2000 seed concepts stochastically selected from the positives and (2) a test set with 1956 candidate concepts comprising the remaining positives and all negatives. This dataset was adopted to explore the feasibility of the proposed WERECE method. In the following experiments, we built a K-means model on seed concepts to verify the proposed approach's effectiveness according to the quality of clustering results. Then, the prediction performance of the proposed discriminant function was compared to that of the K-means model. In essence, based on the strength of the K-means model, we explored the discriminant function's capacity for concept extraction.

The second dataset, EDUTECH-DT, concerned educational technology (i.e., a secondary matter in education) and comprised a collection of seeds along with many candidates. Seeds were drawn from a thesaurus dataset released in our previous research [55], wherein domain experts manually collected educational concepts to improve the standardization of terminology in educational technology. As for candidate concepts, we organized a corpus of academic literature from 2015 to 2020 on educational technology. Each source's title, abstract, and keywords were obtained from the China National Knowledge Internet, one of the most popular academic databases in China. State-of-the-art NLP techniques were used to develop a text preprocessing pipeline comprising word segmentation, stop word removal, part-of-speech (POS) tagging, and duplicate removal. After preprocessing, concept candidates were identified from a list of POS-based linguistic patterns, which could convey the formation and collocation of domain concepts [11,56,57]. Finally, a set of nouns, noun phrases, verbs, and verbal phrases were retained as candidates. Each candidate was further annotated as either positive or negative by two students majoring in educational technology. The Cohen's Kappa score for this annotation step was 0.84, reflecting moderate rater agreement. We referred to the EDUTECH-DT dataset when conducting an initial experiment to explore optimal parameter settings for WERECE. As described in the above section, the proposed method encompassed three key parameters: the dimensions of refined embedding vectors (M), the number of seed concepts (N), and the number of clusters (K). For convenience in implementation, the grid search method was leveraged to determine WERECE's optimal parameters. With optimal parameter settings, the proposed method's effectiveness against baselines was successively verified by means of the bootstrap sampling method. Table 2 lists the basic statistics for our two datasets, including the number of concepts and their length characteristics.

Table 2. Statistics for our experimental datasets.

Statistics		EDU-DT		EDUTECH-DT	
		Training Set	*Test Set*	*Training Set*	*Test Set*
Number of concepts		2000	1956	1832	1016
Concept length	*Max.*	21	18	15	20
	Min.	1	1	1	1
	Average	4.96	4.90	5.11	4.98
	Standard deviation	1.84	2.15	1.70	2.19

4.2. Baselines and Evaluation Metrics

WERECE is a type of unsupervised learning paradigm for domain-specific term extraction. Its utility was tested by taking five well-known unsupervised learning algorithms as baselines: TF-IDF, TextRank, K-means, isolation forest, and one-class support vector machine (SVM). Both TF–IDF and TextRank served as a heuristic to determine candidate concepts' domain relevance; these have been widely used for educational concept extraction [17,19,29]. Different from these two methods, we fit seed concepts to the K-means, one-class SVM, or isolation forest approaches to promote novelty detection. The identified novel terms or phrases could then be deemed educational concepts as suggested in recent studies from multiple fields [29,58–60]. All baselines were trained and tested with our experimental datasets.

WERECE, against the baselines, was evaluated according to precision, recall, and F1 scores, as has been done elsewhere [29,58,61]. Precision represents the ratio of correctly classified domain concepts to all words or phrases that a method holds as domain concepts. Recall is the ratio of correctly classified domain concepts to all domain concepts in a dataset. The F1 score denotes the harmonic tradeoff between precision and recall. These metrics are respectively computed as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{8}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{9}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{10}$$

where TP (i.e., true positive) is the number of domain concepts identified correctly; FP (i.e., false positive) is the number of non-target concepts incorrectly classified as domain concepts; and FN (i.e., false negative) is the number of domain concepts misclassified as non-target concepts.

All methods and evaluation metrics were implemented in Python. In addition to constructing a text preprocessing pipeline, the Chinese NLP toolkit Jieba was used to implement the TF–IDF method. Scikit-learn, a machine learning package in Python, was employed to establish the K-means, one-class SVM, and isolation forest models. The evaluation metrics were developed in scikit-learn as well.

4.3. Experimental Results

This section presents the experimental results and performance comparisons for educational concept extraction.

4.3.1. Results of Feasibility Assessment

With the EDU-DT dataset, the effectiveness of word embedding refinement on K-means clustering was firstly explored; refinement is a prerequisite for clustering in the

WERECE method. Figure 3 depicts the clustering results for the K-means algorithm with or without a re-embedding vector; in this instance, the hyperparameter of the IsoMap algorithm *n_neighbors* was set to 200. The CH score for clustering results upon integrating the K-means algorithm and the re-embedding method was consistently higher than for results obtained with the K-means algorithm alone. This finding suggests that using IsoMap to refine a pre-trained word embedding model can enhance the K-means algorithm's clustering performance. Furthermore, these CH scores declined sharply as the number of clusters increased. Scores later dropped consistently beyond 20 clusters. We also determined that the optimal number of clusters for K-means (with or without word embedding refinement) was about 20 as per the Elbow method.

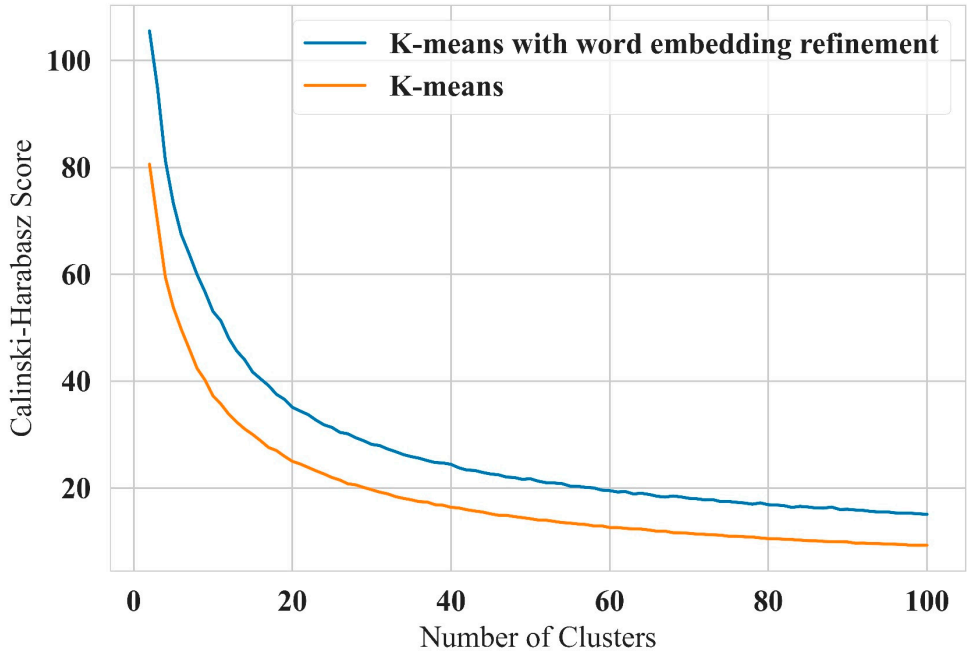

Figure 3. Performance comparison of K-means with and without word embedding refinement on the EDU-DT dataset.

Next, a Mann–Whitney U test was performed to determine whether WERECE's capacity and validity were significantly different from the K-means algorithm in terms of extracting educational concepts. The input for this test included six lists for WERECE and K-means, each containing three sets of 30 scores (precision, recall, and F1). As Figure 4 indicates, WERECE's scores hovered around 0.8, revealing an improvement of approximately 0.1–0.2 versus K-means. WERECE appeared notably capable and valid given the statistically significant performance at the 99% confidence level (all p values were less than 0.01). In particular, the K-means method had a higher standard deviation for each evaluation metric, whereas WERECE scores were much lower. This trend could be attributed to WERECE's Box–Cox transformation; the step diminished the impact from the random selection of original data points on the uncertainty factor associated with clusters yielded by K-means.

The aforementioned findings imply that both the word embedding refinement and discriminant function in the WERECE method are highly feasible and enhance educational concepts' clustering performance. Therefore, concepts were accurately extracted.

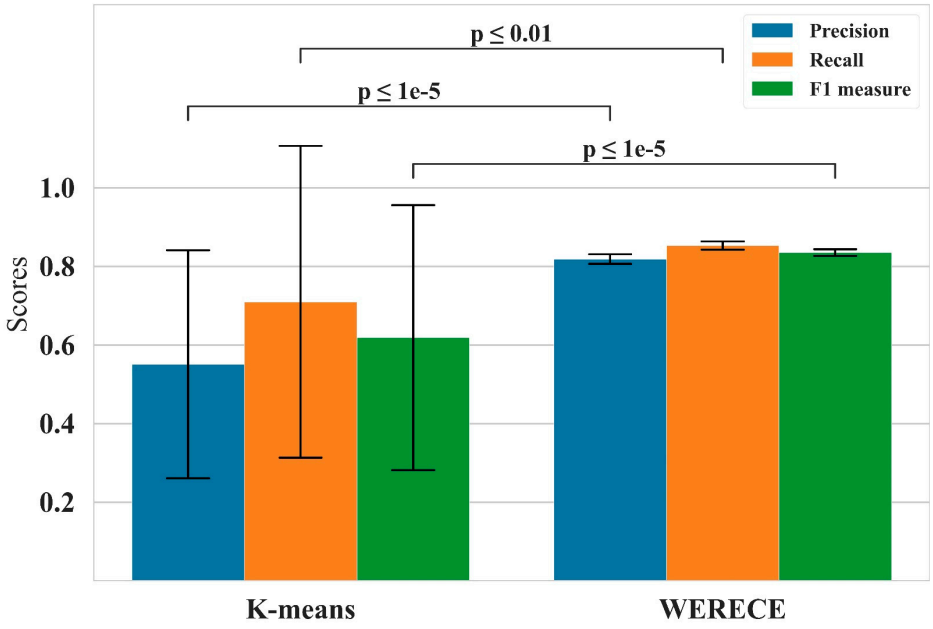

Figure 4. Performance comparisons of WERECE and K-means on the EDU-DT dataset.

4.3.2. Effects of Parameter Settings on WERECE Performance

A series of experiments were completed on the EDUTECH-DT dataset to examine how parameter settings influenced our proposed approach. Figures 5 and 6 showcase how the precision and recall values changed as *M*, *K*, and *N* jointly varied across the pre-specified set of parameter combinations in Table 3.

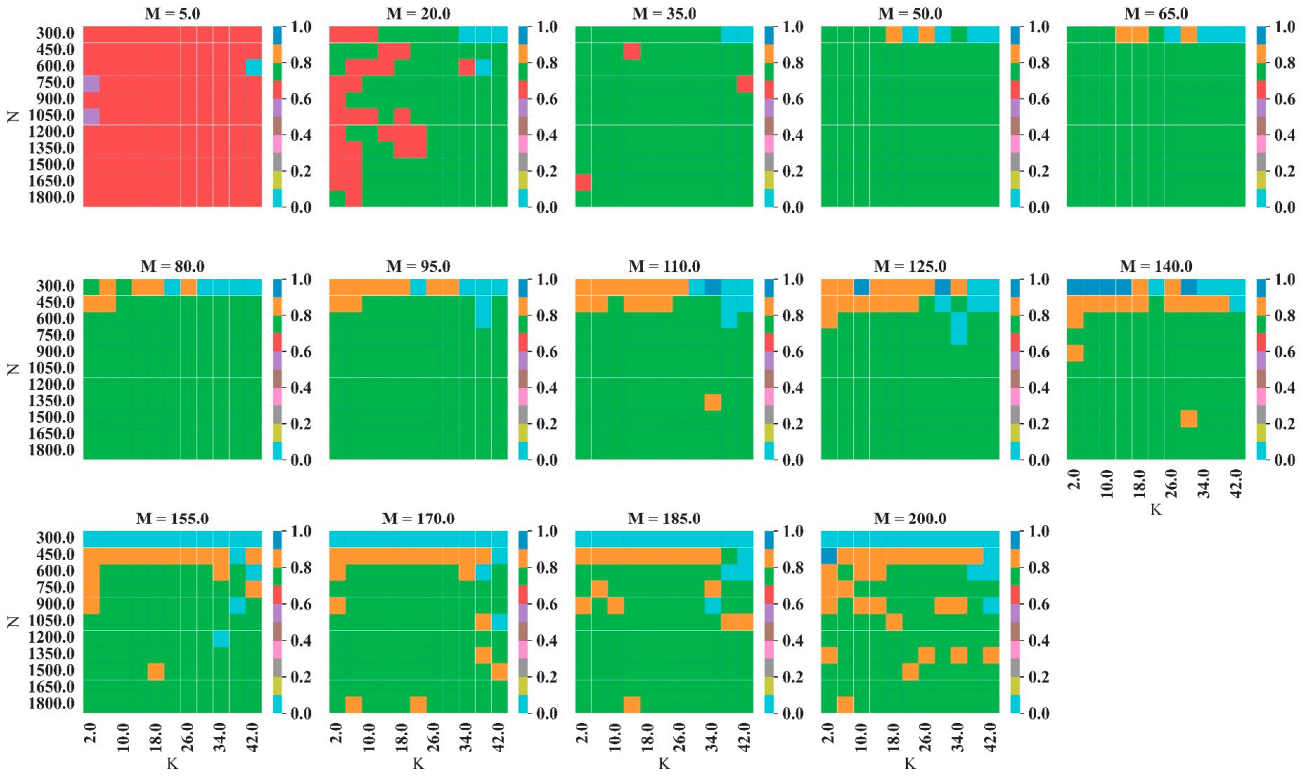

Figure 5. Effects of parameter settings on WERECE's precision. Note that *M* denotes the dimension of refined embedding vectors; *N* represents the number of seeds; *K* is the number of clusters; and the value of precision ranging from 0.0 to 1.0 is colored in sections.

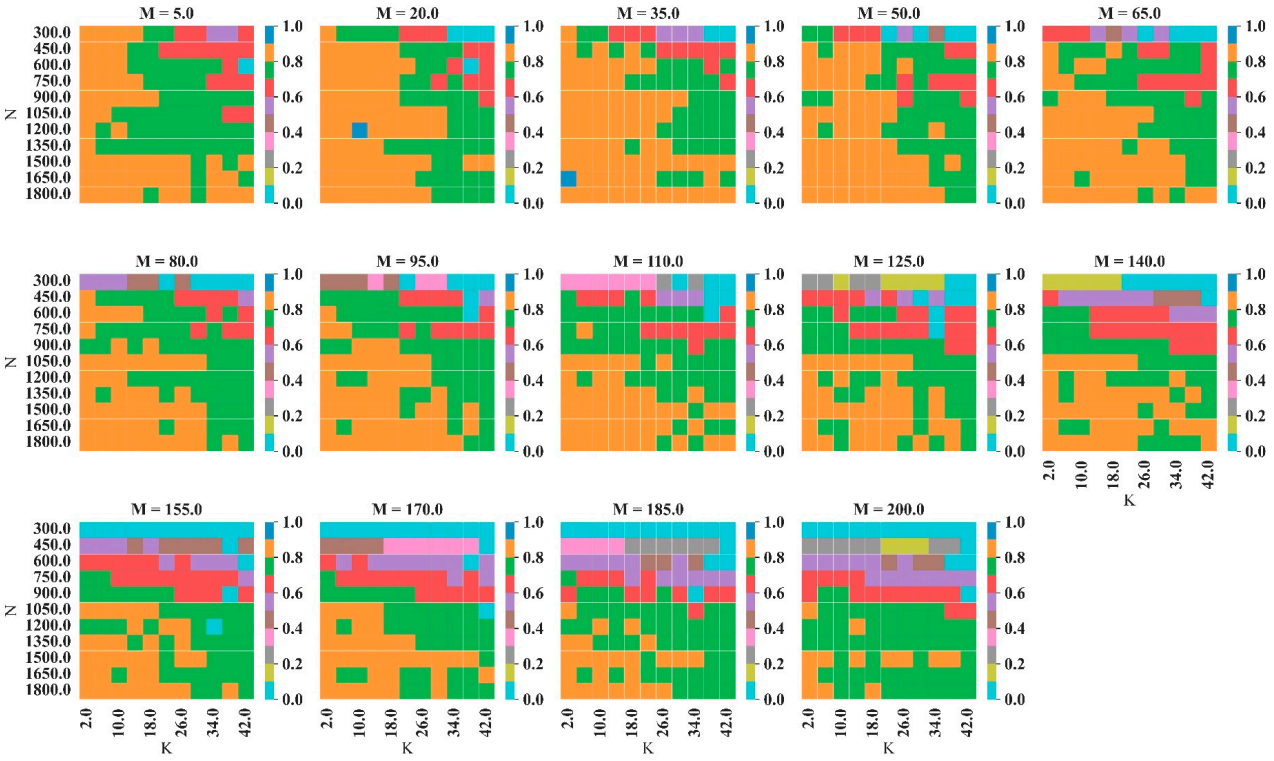

Figure 6. Effects of parameter settings on WERECE's recall. Note that *M* denotes the dimension of refined embedding vectors; *N* represents the number of seeds; *K* is the number of clusters; and the value of recall ranging from 0.0 to 1.0 is colored in sections.

Table 3. Parameter values for the WERECE method.

Parameter	Value Ranges		
	Start	***End***	***Step***
Dimension of refined embedding vectors (*M*)	5	200	15
Number of seeds (*N*)	300	1800	150
Number of clusters (*K*)	2	42	4

As shown in Figure 5, when *M* increased from 5 to 20, WERECE's precision on the test set improved greatly. The proposed method also performed consistently when *M* was between 50 and 140. The re-embedding word embeddings with dimensions above 20 retained substantial semantic information [62]. Additionally, more seeds generally led WERECE to be more precise. The effects of *M* and *K* on precision fell dramatically as *N* rose. For example, WERECE's precision was significantly compromised when *N* was set to 300. This pattern mirrors the argument that insufficient semantic information and highly complex prediction concerns can hamper model performance [63]. Moreover, the number of clusters, *K*, slightly affected WERECE's precision. Thus, when applying a pre-trained embedding model to a domain-specific downstream NLP task, the model must be refined in order to perform well—unless the training dataset contains many samples. In light of these results, it can be found that WERECE demonstrated encouraging performance on educational concept extraction. It was also adequately robust against varied parameter settings (i.e., in terms of precision). An *M* between 50 and 140 appeared optimal. Similarly, the higher the *N*, the better WERECE performed.

Figure 6 illustrates the collective effect of different parameter settings on WERECE's recall performance on the test set. Overall, as *M*, *N*, and *K* jointly increased, the recall scores substantially decreased. Irrespective of the specific values of *N* and *K*, the recall scores exhibited a clear pattern of initially climbing and later falling as *M* rose. This observa-

tion was likely due to redundant semantic information once M exceeded 65. WERECE's recall performance with higher N values declined marginally as M increased; its recall performance with smaller N values decreased greatly. Its recall performance changed more moderately when dealing with smaller (vs. larger) K values. This tendency suggests that WERECE's robustness improved when subjected to a low number of clusters. In particular, taking advantage of a larger N and a smaller K enabled WERECE to mitigate the impact of M variations on recall performance. Hence, WERECE is sensitive to parameter settings in terms of recall and has the potential to produce optimal recall scores when specific parameter settings are met: $N \geq 1000$, $K \leq 18$, and $M \leq 65$.

In summary, by analyzing how parameter settings affected precision and recall, we concluded that WERECE could perform optimally on the EDUTECH-DT dataset when the refined word embedding vector M ranged from 50 to 65, the number of seeds N was greater than 1000, and the number of clusters was smaller than 18.

4.3.3. Comparisons of the Proposed Method with Baselines

To compare the performance of WERECE and baselines, 30 bootstrapped samples were generated by resampling the test set from the EDUTECH-DT dataset. The methods' overall performance on the bootstrapped samples is summarized in Table 4. WERECE with optimal parameter settings ($N = 1000, K = 18, M = 65$) outperformed the baselines in terms of average precision, recall, and F1 scores. TextRank returned the lowest average F1 score, and its recall was significantly higher than its precision ($p < 0.001$) at the 99.9% confidence level. Conversely, one-class SVM performed best among the baselines. WERECE's precision and recall demonstrated comparative improvements (from 0.091 to 0.522 and from 0.078 to 0.16, respectively). The F1 score comprehensively represented precision and recall, resulting in improvements ranging from 0.096 to 0.408. The baselines showed unsatisfactory precision, with the K-means method being particularly unstable. This outcome further confirms that capitalizing on WERECE's discriminant function fortified the method's resilience against outliers. The average recall of the isolation forest was 1.000, implying that this technique is limited in its ability to precisely identify domain concepts. Another experimental study considered this model an ideal choice, given its average recall of 100% [64]. This result might be based on an imbalanced distribution of positive and negative samples or the presence of close similarities between negative and positive samples within a dataset.

Table 4. Performance of WERECE and baselines on the EDUTECH-DT dataset (Mean ± SD).

Method	Evaluation Metrics		
	Precision	**Recall**	**F1 Score**
TextRank	0.337 ± 0.014	0.705 ± 0.028	0.456 ± 0.018
TF–IDF	0.378 ± 0.016	0.792 ± 0.033	0.512 ± 0.022
Isolation Forest	0.544 ± 0.066	1.000 ± 0.000	0.702 ± 0.056
K-means	0.551 ± 0.295	0.710 ± 0.404	0.619 ± 0.343
One-Class SVM	0.768 ± 0.015	0.770 ± 0.037	0.768 ± 0.017
The proposed Method	0.859 ± 0.022	0.870 ± 0.037	0.864 ± 0.023

A Mann–Whitney U test was performed again to determine if WERECE's improvements over the encouraging one-class SVM method were statistically significant. Both methods involved six lists, each with a set of 30 precision, recall, or F1 scores. Figure 7 portrays WERECE's significant improvement over one-class SVM at the 99.9% confidence level ($p < 0.001$). This statistical test reinforced that WERECE significantly outperformed one-class SVM as measured by precision, recall, and F1 scores. The comparison in Table 4 and Figure 7 offers compelling evidence that WERECE can substantially enhance the effectiveness of educational concept extraction.

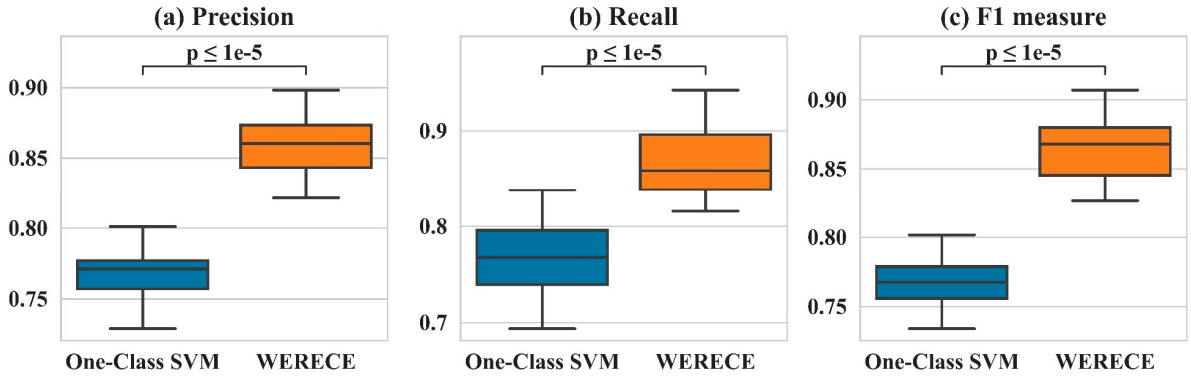

Figure 7. Improvement of WERECE over one-class SVM in terms of evaluation metrics.

5. Discussion

As demonstrated through the experimental results, WERECE exhibits a strong ability in improving the effectiveness of educational concept extraction. A key feature of WERECE lies in that it imposes word re-embedding on a pretrained model to ensure the semantic information of educational concepts are represented precisely. This is consistent with the findings of Albahr et al. [19], who used pre-trained word embeddings derived from Wikipedia to outperform TF-IDF on educational concept extraction. Also, a manifold learning algorithm adopted in WERECE allows reuse of ready-made pretrained models, with the aim of saving time on training and improving the performance of educational concept extraction. This case highlights the importance of manifold learning in discovering the potential geometric relationships among concept representations in a high-dimensional semantic space. According to Wang et al. [65], manifold learning-based re-embedding ensures the consistency of the geometric relationship among concept representations and the real semantic information among concepts, thereby improving the performance of downstream NLP tasks. Furthermore, the experimental results demonstrated that the Box–Cox transformation consistently boosts the effectiveness of WERECE's discriminant function. The Box–Cox transformation was conducive in mitigating the uncertainty factor of K-means clusters caused by the random selection of original data points. Such an aspect ties in well with previous studies showing the impact of the Box–Cox transformation on machine learning algorithms [66,67].

WERECE achieved the best performance on educational concept extraction in the balance of precision and recall when compared to baselines. In light of our experimental findings, it is evident that both TF-IDF and TextRank present a deficiency in precision. This is attributed to their inadequate extraction of certain frequent candidate concepts as educational concepts while simultaneously disregarding infrequent educational concepts. This also accords with the findings of a great deal of the previous work in educational concept extraction [17,19,29,61,68]. For example, TF-IDF achieved an average precision of approximately 0.35 in educational concept extraction from MOOC (i.e., Massive Open Online Course) video lectures [19]; the average precision of educational concept extraction through TextRank is around 0.45 [29]. It was suggested that isolation forest shows the best effectiveness results in domain concept extraction [59]. However, based on the empirical evidence obtained from our experiments, this does not appear to be the case. Although isolation forest achieved the highest recall in education concept extraction, its precision and F1 score are unsatisfactory. A possible explanation for this inconsistency might be that isolation forest is sensitive to sample distribution in experimental datasets [64]. With regard to K-means and one-class SVM in our experiments, their performance broadly supports the work of other studies in education, information retrieval, and medical services. In E-learning systems, K-means can be employed to obtain domain concept extraction from scholarly articles; it achieved an average precision of 0.80 [69]. To construct a target-specific sentiment lexicon, one-class SVM yielded a decent outcome, demonstrating an average precision of 0.79 and an average recall of 0.75 [58]. In the field of medicine, K-means was

used to produce semantic clusters to identify medical terms, which facilitates verification of the impact of medical terms used by doctors on the service quality of E-health [60]. Compared to baselines, the encouraging precision and recall of WERECE measured in our experiments strengthen the case for bringing semantic association and computation into domain concept extraction. In addition, the introduction of word embedding refinement and the Box–Cox transformation is a recipe for the trade-off between WERECE's precision and recall. Thus, WERECE is an effective method and provides promising potential for educational concept extraction.

Although the strengths of WERECE are encouraging, several limitations provide intriguing directions for later work. The experimental datasets revolved around education and educational technology; WERECE must therefore be tested in diverse disciplines to affirm its generalizability. We believe that optimizing experimental datasets will enhance the model's performance in terms of concept extraction. Particularly, the construction of benchmark datasets allows fair comparison of the performance between WERECE and baselines. Meanwhile, this study was carried out in a monolingual setting. A language-independent method for educational concept extraction can yield interconnected concepts in multilingual contexts [70]. We also investigated how different parameter settings affected WERECE's performance but did not certify its optimal parameter settings. Multi-objective hyperparameter optimization algorithms (e.g., evolution strategies and genetic algorithms) may address this issue [71]. In addition, to simplify WERECE's implementation, candidate concepts were generated based on lexical-syntactic patterns. Including other filtering criteria to produce candidates would further improve the proposed method's accuracy. Necessarily study of these limitations is thus required for the improvement of WERECE's accuracy and robustness in domain adaption.

6. Conclusions and Future Work

A vast volume of educational data has emerged out of massive online learning and new learning media. More useful information must be extracted from educational big data to inform artificial intelligence–empowered teaching and learning. We intended to exploit semantic information for educational concept extraction (i.e., knowledge mining and information processing). Hence WERECE was proposed, a simple yet powerful unsupervised educational concept extraction method based on word embedding refinement. To the best of our knowledge, this effort represents an early attempt to refine pre-trained word embeddings for educational concept extraction in an unsupervised manner. Various experiments on two educational datasets substantiated WERECE's feasibility and effectiveness. We also observed how its parameter settings influenced the model's performance. Experimental results related to educational concept extraction demonstrate promise in three regards. First, two core components of WERECE—word embedding refinement and the discriminant function—could enrich educational concept extraction. Second, WERECE showed robust precision and sensitive recall, even with varied parameter settings. Third, it significantly outperformed the chosen baselines with respect to educational concept extraction. As a conclusion, WERECE provides valuable insight into the integration of pretrained word embedding models in the unsupervised extraction of educational concepts.

In the future, we intend to apply WERECE to corpora beyond education and educational technology. Genetic algorithms can be used to identify the model's optimal parameter settings to achieve the best performance. WERECE's output could also act as a springboard for further educational applications. For instance, the model can be used to derive discipline-specific concepts from students' online discussions. The concepts' semantic similarity can then be computed to uncover their semantic relationships (e.g., subordinate relations). Such work can inform a discipline-oriented knowledge graph suitable for various educational tasks, such as learning assessment and feedback, learning interest prediction, and learning resource recommendation.

Author Contributions: Conceptualization, Y.Z. and J.H.; methodology, J.H. and Y.Z.; software, R.D.; validation, R.D. and X.W.; formal analysis, Y.Z. and J.H.; investigation, S.C., R.D., J.Z. and L.L.; resources, X.W., S.C., J.Z. and L.L.; data curation, S.C.; writing—original draft preparation, J.H., R.D. and X.W.; visualization, J.H.; supervision, Y.Z.; project administration, Y.Z.; funding acquisition, J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China: 62207015, and the Humanities and Social Sciences Youth Foundation of the Chinese Ministry of Education: 22YJC880021.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

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

References

1. Bai, X.; Zhang, F.; Li, J.; Guo, T.; Aziz, A.; Jin, A.; Xia, F. Educational Big Data: Predictions, Applications and Challenges. *Big Data Res.* **2021**, *26*, 100270. [CrossRef]
2. Shamsfard, M.; Barforoush, A.A. The State of the Art in Ontology Learning: A Framework for Comparison. *Knowl. Eng. Rev.* **2003**, *18*, 293–316. [CrossRef]
3. Poria, S.; Hussain, A.; Cambria, E.; Poria, S.; Hussain, A.; Cambria, E. Concept Extraction from Natural Text for Concept Level Text Analysis. In *Multimodal Sentiment Analysis*; Springer: Cham, Switzerland, 2018; pp. 79–84.
4. Fu, S.; Chen, D.; He, H.; Liu, S.; Moon, S.; Peterson, K.J.; Shen, F.; Wang, L.; Wang, Y.; Wen, A.; et al. Clinical Concept Extraction: A Methodology Review. *J. Biomed. Inform.* **2020**, *109*, 103526. [CrossRef] [PubMed]
5. Firoozeh, N.; Nazarenko, A.; Alizon, F.; Daille, B. Keyword Extraction: Issues and Methods. *Nat. Lang. Eng.* **2020**, *26*, 259–291. [CrossRef]
6. Szwed, P. Concepts Extraction from Unstructured Polish Texts: A Rule Based Approach. In Proceedings of the 2015 Federated Conference on Computer Science and Information Systems (FedCSIS), Lodz, Poland, 13–16 September 2015; pp. 355–364.
7. Stanković, R.; Krstev, C.; Obradović, I.; Lazić, B.; Trtovac, A. Rule-Based Automatic Multi-Word Term Extraction and Lemmatization. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 23–28 May 2016; pp. 507–514.
8. Gong, L.; Yang, R.; Liu, Q.; Dong, Z.; Chen, H.; Yang, G. A Dictionary-Based Approach for Identifying Biomedical Concepts. *Int. J. Pattern Recognit. Artif. Intell.* **2017**, *31*, 1757004. [CrossRef]
9. Aizawa, A. An Information-Theoretic Perspective of Tf--Idf Measures. *Inf. Process. Manag.* **2003**, *39*, 45–65. [CrossRef]
10. Mihalcea, R.; Tarau, P. Textrank: Bringing Order into Text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 25–26 July 2004; pp. 404–411.
11. Zhang, Z.; Gao, J.; Ciravegna, F. Semre-Rank: Improving Automatic Term Extraction by Incorporating Semantic Relatedness with Personalised Pagerank. *ACM Trans. Knowl. Discov. Data* **2018**, *12*, 1–41. [CrossRef]
12. Tulkens, S.; Šuster, S.; Daelemans, W. Unsupervised Concept Extraction from Clinical Text through Semantic Composition. *J. Biomed. Inform.* **2019**, *91*, 103120. [CrossRef]
13. Xiong, A.; Liu, D.; Tian, H.; Liu, Z.; Yu, P.; Kadoch, M. News Keyword Extraction Algorithm Based on Semantic Clustering and Word Graph Model. *Tsinghua Sci. Technol.* **2021**, *26*, 886–893. [CrossRef]
14. Daems, O.; Erkens, M.; Malzahn, N.; Hoppe, H.U. Using Content Analysis and Domain Ontologies to Check Learners' Understanding of Science Concepts. *J. Comput. Educ.* **2014**, *1*, 113–131. [CrossRef]
15. Abyaa, A.; Khalidi Idrissi, M.; Bennani, S. Learner Modelling: Systematic Review of the Literature from the Last 5 Years. *Educ. Technol. Res. Dev.* **2019**, *67*, 1105–1143. [CrossRef]
16. Chen, N.-S.; Wei, C.-W.; Chen, H.-J. Mining E-Learning Domain Concept Map from Academic Articles. *Comput. Educ.* **2008**, *50*, 1009–1021. [CrossRef]
17. Conde, A.; Larrañaga, M.; Arruarte, A.; Elorriaga, J.A.; Roth, D. Litewi: A Combined Term Extraction and Entity Linking Method for Eliciting Educational Ontologies from Textbooks. *J. Assoc. Inf. Sci. Technol.* **2016**, *67*, 380–399. [CrossRef]
18. Pan, L.; Wang, X.; Li, C.; Li, J.; Tang, J. Course Concept Extraction in MOOCS via Embedding-Based Graph Propagation. In Proceedings of the Eighth International Joint Conference on Natural Language Processing, Taipei, Taiwan, 28–30 November 2017; Asian Federation of Natural Language Processing: Volume 1: Long Papers; pp. 875–884.
19. Albahr, A.; Che, D.; Albahar, M. A Novel Cluster-Based Approach for Keyphrase Extraction from MOOC Video Lectures. *Knowl. Inf. Syst.* **2021**, *63*, 1663–1686. [CrossRef]

20. Liu, J.; Shao, X. An Improved Extracting Chinese Term Method Based on C/NC-Value. In Proceedings of the 2010 International Symposium on Intelligence Information Processing and Trusted Computing, Wuhan, China, 28–29 October 2010; IEEE: Piscataway, NJ, USA, 2010.
21. Zhang, L.; Li, X.-P.; Zhang, F.-B.; Hu, B. Research on Keyword Extraction and Sentiment Orientation Analysis of Educational Texts. *J. Comput.* **2017**, *28*, 301–313.
22. Lu, M.; Wang, Y.; Yu, J.; Du, Y.; Hou, L.; Li, J. Distantly Supervised Course Concept Extraction in MOOCs with Academic Discipline. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; Volume 1: Long Papers.
23. Kang, Y.-B.; Haghighi, P.D.; Burstein, F. CFinder: An Intelligent Key Concept Finder from Text for Ontology Development. *Expert Syst. Appl.* **2014**, *41*, 4494–4504. [CrossRef]
24. Levow, G.-A.; Oard, D.W.; Resnik, P. Dictionary-Based Techniques for Cross-Language Information Retrieval. *Inf. Process. Manag.* **2005**, *41*, 523–547. [CrossRef]
25. Bellaachia, A.; Al-Dhelaan, M. NE-Rank: A Novel Graph-Based Keyphrase Extraction in Twitter. In Proceedings of the 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Macau, China, 4–7 December 2012; IEEE: Piscataway, NJ, USA, 2012.
26. Bougouin, A.; Boudin, F.; Daille, B. Topicrank: Graph-Based Topic Ranking for Keyphrase Extraction. In Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), Nagoya, Japan, 14–19 October 2013; pp. 543–551.
27. Boudin, F. Unsupervised Keyphrase Extraction with Multipartite Graphs. *arXiv* **2018**, arXiv:1803.08721.
28. Kong, S.C.; Li, P.; Song, Y. Evaluating a Bilingual Text-Mining System with a Taxonomy of Key Words and Hierarchical Visualization for Understanding Learner-Generated Text. *ACM J. Educ. Resour. Comput.* **2018**, *56*, 369–395. [CrossRef]
29. Chau, H.; Labutov, I.; Thaker, K.; He, D.; Brusilovsky, P. Automatic Concept Extraction for Domain and Student Modeling in Adaptive Textbooks. *Int. J. Artif. Intell. Educ.* **2021**, *31*, 820–846. [CrossRef]
30. Peng, X.; Han, C.; Ouyang, F.; Liu, Z. Topic Tracking Model for Analyzing Student-Generated Posts in SPOC Discussion Forums. *Int. J. Educ. Technol. High. Educ.* **2020**, *17*, 35. [CrossRef]
31. 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; Association for Computational Linguistics: Stroudsburg, PA, USA, 2013; pp. 746–751.
32. Niven, T.; Kao, H.-Y. Probing Neural Network Comprehension of Natural Language Arguments. *arXiv* **2019**, arXiv:1907.07355.
33. Chalkidis, I.; Fergadiotis, M.; Malakasiotis, P.; Aletras, N.; Androutsopoulos, I. LEGAL-BERT: The Muppets Straight out of Law School. *arXiv* **2020**, arXiv:2010.02559.
34. Wang, Y.; Liu, S.; Afzal, N.; Rastegar-Mojarad, M.; Wang, L.; Shen, F.; Kingsbury, P.; Liu, H. A Comparison of Word Embeddings for the Biomedical Natural Language Processing. *J. Biomed. Inform.* **2018**, *87*, 12–20. [CrossRef] [PubMed]
35. Clavié, B.; Gal, K. Edubert: Pretrained Deep Language Models for Learning Analytics. *arXiv* **2019**, arXiv:1912.00690.
36. Sezerer, E.; Tekir, S. A Survey on Neural Word Embeddings. *arXiv* **2021**, arXiv:2110.01804.
37. Wang, S.; Zhang, Y.; Shi, W.; Zhang, G.; Zhang, J.; Lin, N.; Zong, C. A Large Dataset of Semantic Ratings and Its Computational Extension. *Sci Data* **2023**, *10*, 106. [CrossRef]
38. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
39. Song, Y.; Shi, S.; Li, J.; Zhang, H. Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; Volume 2 (Short Papers).
40. Patel, A.; Sands, A.; Callison-Burch, C.; Apidianaki, M. Magnitude: A Fast, Efficient Universal Vector Embedding Utility Package. *arXiv* **2018**, arXiv:1810.11190.
41. Hasan, S.; Curry, E. Word Re-Embedding via Manifold Dimensionality Retention. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017.
42. Yonghe, C.; Lin, H.; Yang, L.; Diao, Y.; Zhang, S.; Xiaochao, F. Refining Word Reesprentations by Manifold Learning. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019; pp. 5394–5400.
43. Zhao, W.; Zhou, D.; Li, L.; Chen, J. Manifold Learning-Based Word Representation Refinement Incorporating Global and Local Information. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; International Committee on Computational Linguistics: Stroudsburg, PA, USA, 2020.
44. Tenenbaum, J.B.; Silva, V.D.; Langford, J.C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **2000**, *290*, 2319–2323. [CrossRef]
45. Xu, D.; Tian, Y. A Comprehensive Survey of Clustering Algorithms. *Ann. Data Sci.* **2015**, *2*, 165–193. [CrossRef]
46. Calinski, T.; Harabasz, J. A Dendrite Method for Cluster Analysis. *Commun. Stat. Theory Methods* **1974**, *3*, 1–27. [CrossRef]
47. Berkhin, P. A Survey of Clustering Data Mining Techniques. In *Grouping Multidimensional Data*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 25–71, ISBN 9783540283485.
48. Désir, C.; Bernard, S.; Petitjean, C.; Heutte, L. One Class Random Forests. *Pattern Recognit.* **2013**, *46*, 3490–3506. [CrossRef]

49. Box, G.E.P.; Cox, D.R. An Analysis of Transformations. *J. R. Stat. Soc.* **1964**, *26*, 211–243. [CrossRef]
50. Utsumi, A. Exploring What Is Encoded in Distributional Word Vectors: A Neurobiologically Motivated Analysis. *Cogn. Sci.* **2020**, *44*, e12844. [CrossRef] [PubMed]
51. Peterson, K.J.; Liu, H. An Examination of the Statistical Laws of Semantic Change in Clinical Notes. *AMIA Jt. Summits Transl. Sci. Proc.* **2021**, *2021*, 515–524.
52. Magister, L.C.; Barbiero, P.; Kazhdan, D.; Siciliano, F.; Ciravegna, G.; Silvestri, F.; Jamnik, M.; Lio, P. Encoding Concepts in Graph Neural Networks. *arXiv* **2022**, arXiv:2207.13586.
53. Yu, J.; Luo, G.; Xiao, T.; Zhong, Q.; Wang, Y.; Feng, W.; Luo, J.; Wang, C.; Hou, L.; Li, J.; et al. MOOCCube: A Large-Scale Data Repository for NLP Applications in MOOCs. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 3135–3142.
54. Lin, Y.; Feng, S.; Lin, F.; Zeng, W.; Liu, Y.; Wu, P. Adaptive Course Recommendation in MOOCs. *Knowl.-Based Syst.* **2021**, *224*, 107085. [CrossRef]
55. Wu, L.; Liu, Q.; Zhao, G.; Huang, H.; Huang, T. Thesaurus Dataset of Educational Technology in Chinese. *Br. J. Educ. Technol.* **2015**, *46*, 1118–1122. [CrossRef]
56. Kang, Y.-B.; Haghigh, P.D.; Burstein, F. TaxoFinder: A Graph-Based Approach for Taxonomy Learning. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 524–536. [CrossRef]
57. Desul, S.; Madurai Meenachi, N.; Venkatesh, T.; Gunta, V.; Gowtham, R.; Magapu, S.B. Method for Automatic Key Concepts Extraction: Application to Documents in the Domain of Nuclear Reactors. *Electron. Libr.* **2018**, *37*, 2–15. [CrossRef]
58. Wu, S.; Wu, F.; Chang, Y.; Wu, C.; Huang, Y. Automatic Construction of Target-Specific Sentiment Lexicon. *Expert Syst. Appl.* **2019**, *116*, 285–298. [CrossRef]
59. Papagiannopoulou, E.; Tsoumakas, G.; Papadopoulos, A. Keyword Extraction Using Unsupervised Learning on the Document's Adjacency Matrix. In Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15), Mexico City, Mexico, 11 June 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021.
60. Zhang, J.; Zhang, J.; Wang, K.; Yan, W. Should Doctors Use or Avoid Medical Terms? The Influence of Medical Terms on Service Quality of E-Health. *Electr. Commer. Res.* **2021**, *23*, 1775–1805. [CrossRef]
61. Lu, W.; Zhou, Y.; Yu, J.; Jia, C. Concept Extraction and Prerequisite Relation Learning from Educational Data. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 9678–9685.
62. Zhao, D.; Wang, J.; Chu, Y.; Zhang, Y.; Yang, Z.; Lin, H. Improving Biomedical Word Representation with Locally Linear Embedding. *Neurocomputing* **2021**, *447*, 172–182. [CrossRef]
63. Sharif, W.; Mumtaz, S.; Shafiq, Z.; Riaz, O.; Ali, T.; Husnain, M.; Choi, G.S. An Empirical Approach for Extreme Behavior Identification through Tweets Using Machine Learning. *Appl. Sci.* **2019**, *9*, 3723. [CrossRef]
64. Aminanto, M.E.; Ban, T.; Isawa, R.; Takahashi, T.; Inoue, D. Threat Alert Prioritization Using Isolation Forest and Stacked Auto Encoder with Day-Forward-Chaining Analysis. *IEEE Access* **2020**, *8*, 217977–217986. [CrossRef]
65. Wang, B.; Sun, Y.; Chu, Y.; Lin, H.; Zhao, D.; Yang, L.; Shen, C.; Yang, Z.; Wang, J. Manifold biomedical text sentence embedding. *Neurocomputing* **2022**, *492*, 117–125. [CrossRef]
66. Bicego, M.; Baldo, S. Properties of the Box–Cox transformation for pattern classification. *Neurocomputing* **2016**, *218*, 390–400. [CrossRef]
67. Blum, L.; Elgendi, M.; Menon, C. Impact of Box-Cox Transformation on Machine-Learning Algorithms. *Front. Artif. Intell.* **2022**, *5*, 877569. [CrossRef]
68. Wang, X.; Feng, W.; Tang, J.; Zhong, Q. Course concept extraction in MOOC via explicit/implicit representation. In Proceedings of the Third International Conference on Data Science in Cyberspace (DSC), Guangzhou, China, 18–21 June 2018; IEEE: Piscataway, NJ, USA, 2018.
69. Ahmed, R.; Ahmad, T.; Almutairi, F.M.; Qahtani, A.M.; Alsufyani, A.; Almutiry, O. Fuzzy semantic classification of multi-domain E-learning concept. *Mob. Netw. Appl.* **2021**, *26*, 2206–2215. [CrossRef]
70. Alba, A.; Coden, A.; Gentile, A.L.; Gruhl, D.; Ristoski, P.; Welch, S. Multi-Lingual Concept Extraction with Linked Data and Human-in-the-Loop. In Proceedings of the Knowledge Capture Conference, Austin, TX, USA, 4–6 December 2017; ACM: New York, NY, USA, 2017.
71. Morales-Hernández, A.; Van Nieuwenhuyse, I.; Rojas Gonzalez, S. A Survey on Multi-Objective Hyperparameter Optimization Algorithms for Machine Learning. *Artif. Intell. Rev.* **2023**, *56*, 8043–8093. [CrossRef]