SCIENTIFIC Publishing Limited

Digital Technologies Research and Applications

http://ojs.ukscip.com/index.php/dtra

Article

Evaluating Semantic Representation Strategies for Robust Information Retrieval Matching

Eoin O Connell [1], **Niall McCarroll** [2], **Sujata Rani** [2], **Kevin Curran** [2,*], **Eugene McNamee** [3], **Angela Clist** [1] and **Andrew Brammer** [1]

[1] A&O Shearman, 68 Donegall Quay, Belfast BT1 3NL, Northern Ireland

[2] School of Computing, Engineering and Intelligent Systems, Ulster University, Derry BT48 7JL, Northern Ireland

[3] School of Law, Ulster University, Belfast BT15 1AP, Northern Ireland

[*] Correspondence: kj.curran@ulster.ac.uk

Received: 20 August 2025; **Revised:** 3 September 2025; **Accepted:** 26 September 2025; **Published:** 11 October 2025

Abstract: Vector Space Models (VSM) and neural word embeddings are core components in recent Machine Learning (ML) and Natural Language Processing (NLP) pipelines. By encoding words, sentences and documents as high-dimensional vectors via distributional semantics, they enable Information Retrieval (IR) systems to capture semantic relatedness between queries and answers. This paper compares different semantic representation strategies for query-statement matching, evaluating paraphrase identification within an IR framework using partial and syntactically varied queries of different lengths. Motivated by the Word Mover's Distance (WMD) model, similarity is evaluated using the distance between individual words of queries and statements, as opposed to the common similarity measure of centroids of neural word embeddings. Results from ranked query and response statements demonstrate significant gains in accuracy using the combined approach of similarity ranking through WMD with the word embedding techniques. Our top-performing WMD + *GloVe* system consistently outperformed *Doc2Vec* and an LSA baseline across three return-rate thresholds, achieving 100% correct matches within the top-3 ranked results and 89.83% top-1 accuracy. Beyond the substantial gains from WMD-based similarity ranking, our results indicate that large, pre-trained word embeddings, trained on vast amounts of data, result in portable, domain-agnostic language processing solutions suitable for diverse business use cases.

Keywords: Semantic Information Retrieval; Word Embeddings; Document Similarity; Query-Statement Matching; *GloVe*; WMD

1. Introduction

Information retrieval (IR) can be considered from various levels of processing, including single words or sentences to paragraphs and full document retrieval. The most significant challenge is retrieving documents that are relevant to user queries. The ability to rank results by query relevance is a key aspect of Information Retrieval, and it is this function that differentiates IR from other types of database queries, which are sorted by one or more table columns. IR systems, such as search engines, return results sorted in descending order based on a score that designates the strength of the match between the query and the returned document [1]. If we consider business scenarios where potentially millions of documents need to be searched and processed, ranked retrieval has significant implications for efficiency, as it prevents users from being overloaded with results that are impossible to

navigate and consume [2]. The concept of synonyms or semantic heterogeneity in language is one in which the same real-world entity can be represented using different linguistic terms. For example, the concept of the word 'beautiful' can also be conveyed with the words 'attractive,' 'pretty,' 'lovely,' and 'stunning.' The ability to manage synonymy has important implications for querying data from multiple sources, as well as the cleansing and mining of data [3–5]. Polysemy is the capacity of a word or phrase to have multiple meanings. For example, the word 'glass' has two different meanings when we compare its use in the sentence: 'I emptied the glass (container)' to its use in the sentence 'I drank the glass (the contents of the container)' [6].

Many traditional keyword-based information retrieval systems rely on statistical term overlap. This direct mapping of a query with indexed terms or statements suffers from lexical gaps when considering Paraphrase Identification (PI), where conceptually identical or similar statements can be expressed in various ways [7]. Conversely, it is also worth noting that completely unrelated concepts can be textually quite similar. As a result, these traditional information retrieval systems often miss relevant documents or return irrelevant ones that contain different terms than the query, even with the use of query expansion [8]. The ability to identify related terms outside of the keyword range is also important for handling partial queries or scenarios where users conduct exploratory searches based on minimal or imprecise details [9]. These complexities have driven the development of Natural Language Processing (NLP) beyond pure text-based search solutions, enabling the interpretation of language in a more complex and meaningful way [10, 11]. Semantic analysis extends beyond word-to-object associations to uncover the links between the broader set of words that can be attributed to each object.

Several machine learning (ML) and statistical strategies can be employed to estimate and uncover the underlying latent structure of meaning. These algorithms work to organize text into a semantic structure that can be leveraged to maximize the representation and retrieval of information, thus facilitating information navigation [12]. Semantic similarity matching has been shown to improve recall and precision and has a wide range of applications within NLP, including plagiarism detection [13], text summarization [14, 15], evaluation of text coherence [16, 17], word sense disambiguation [18], text categorization, relevance feedback [19,20], and sentiment analysis [21]. Semantic learning is also considered one of the most effective techniques for improving the effectiveness of information retrieval [22]. Word and paragraph embedding algorithms such as *Word2Vec* [23], *Doc2Vec* [24], *GloVe* [25], and *FastText* [26], have emerged as leading approaches to modelling the semantic relations between terms in various NLP pipelines. The semantic-based IR process can be considered an implementation of two phases. The first stage involves processing text into a semantic vector space. The second stage involves ranking of candidates through a mechanism of similarity comparison. The similarity technique used in this paper is the Word Mover's distance (WMD), which measures query-statement similarity based on an evaluation of the distance between individual words [27], as opposed to the common similarity measure that uses query-statement centroids of word embeddings.

This paper evaluates existing and newly proposed models that integrate these pre-trained neural word vector embeddings into the matching and ranking phases of the information retrieval pipeline. The main research question addressed in this paper is how to perform information retrieval by considering the semantics of query-statement matching using neural word embeddings. In addressing this, the research makes the following important contributions.

Contributions

1. The systematic evaluation of WMD combined with *Word2Vec, FastText*, and *GloVe* embeddings for semantic information retrieval, compared with LSA and *Doc2Vec* baselines.
2. Demonstrated robust retrieval performance on paraphrased and partial query matches, showing the ability to capture semantic variation.
3. Investigated retrieval robustness across query lengths, from short statements to multi-sentence paragraphs.
4. Established WMD + *GloVe* as the most effective approach, achieving superior ranking accuracy, particularly in top-ranked positions.

The remainder of this paper is organized as follows: Section 2 discusses related work and reviews existing embedding-based approaches for conceptual search strategies before highlighting the research objective of this study. Section 3 details the methodology and experimental setup that was implemented to evaluate the different word embedding models on a practical information retrieval task. Section 4 discusses results and evaluation of the

query-statement matching trials before concluding with a detailed discussion of findings in Section 5 and concluding remarks in Section 6. Section 7 discusses the limitations of the system and future work of the study.

2. Background and Related Work

The challenge of matching documents or statements based on their textual descriptions remains an active area of research in Information Retrieval. Moving away from the traditional approach of counting query term occurrences in the target search text, many latent semantic and more recent neural embedding methods have been proposed to bridge the gap caused by linguistic and vocabulary-related mismatches and differences. Algorithmic relevance is at the computational core of Information Retrieval and concerns the relationship between information objects and user queries based on some measure of similarity between them. The 'gold standard' of algorithmic relevance performance is that the search engine query should retrieve specified sets of information objects, measured as recall with a minimum number of false positives, measured as precision [22].

2.1. Background

2.1.1. Traditional Vector-Based Algorithms

Traditional vector-based algorithms such as Okapi BM25 and TF-IDF (Term Frequency – Inverse Document Frequency) count term occurrences and utilize bag-of-words representations reweighted by inverse document frequency [28]. In the bag-of-words approach, text documents are represented by isolated keyword terms that have no syntactic or semantic context or relation to other terms in the model. While these approaches are strong baselines, they fail to adequately represent complex text objects such as sentences and paragraphs because all relationships and term dependencies are lost.

The use of latent semantic structures to facilitate Information Retrieval is well established [29]. Semantic analysis models [30], such as Latent Semantic Analysis (LSA) or Latent Semantic Indexing (LSI) map dense vector representations onto a low-dimensional subspace for corpus-based similarity comparisons [31]. Since LSA calculates similarity based on context, it does not require queries, target statements, or documents to contain common words; therefore, it outperforms traditional vector space models in measures of synonymy. However, as a global bag-of-words approach, it fails to efficiently leverage lower-level syntactic and statistical information that exposes the links between component vectors. As a consequence, LSA tends to be more appropriate for similarity matching of longer texts as opposed to keyword matching [32].

2.1.2. Embeddings

The move to bridge the lexical gap caused by linguistic differences and to adequately respond to the challenge of representing documents semantically has prompted the development of advanced representation techniques, such as Distributional Semantic Models (DSMs), that signify a move away from simple syntactic matching mechanisms to more complex combinations of syntactic and semantic parsing that enhance search recall and expressiveness [33]. Computational linguistics has consistently demonstrated that contextual information provides a reliable approximation of word meaning, as semantically related words tend to occur in similar contextual distributions [34,35]. DSMs employ vectors to track these contexts in which target terms appear and store them as meaning representations. Geometric techniques are then applied to these vectors to measure the similarity in meaning between search and target phrases [35].

From the statistical neural net, language models evolved word embeddings learned by neural networks [36]. These word embeddings learn semantic word vectors to predict context words, thus capturing both the syntactic and semantic relations between the collections of words that constitute a sentence or paragraph [25,37].

Word2Vec: Mikolov et al. [23] proposed *Word2Vec*, an unsupervised, shallow neural network-based skip-gram model in which word vector representations are learned by reconstructing each word's context through an efficient training algorithm that does not utilize dense matrix manipulation (**Figure 1**). Recall is significantly enhanced over keyword matching techniques as *Word2Vec* works to bring semantically synonymous vectors closer together in the embedded space. The original implementation of *Word2Vec* uses centroids of word vectors and cosine similarity to evaluate document relatedness.

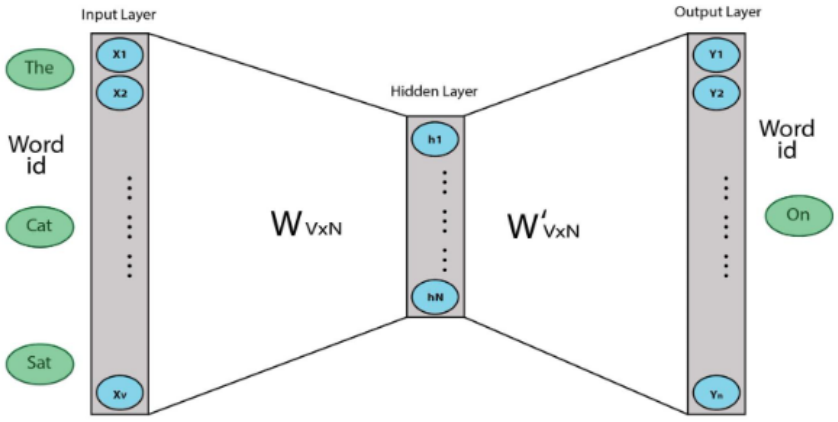

Figure 1. The *Word2Vec* neural network architecture.

Doc2Vec: Taking the concept of context beyond single words, Le and Mikolov [24] further extended the *Word2Vec* framework with a semantic enriching strategy that learns fixed-length feature representations from variable-length segments of text, such as sentences, paragraphs, and entire documents. *Doc2Vec* (or paragraph vectors) represents a combined approach where each word and each paragraph are mapped to unique vectors (known as word embeddings and paragraph embeddings, respectively) (**Figure 2**). Their experiments found that the *Doc2Vec* framework, with its enhanced decision-making capability, proved successful in information retrieval and sentiment analysis tasks, reporting less retrieval and classification errors than comparable algorithms that employed Word Centroid Distance similarity measures.

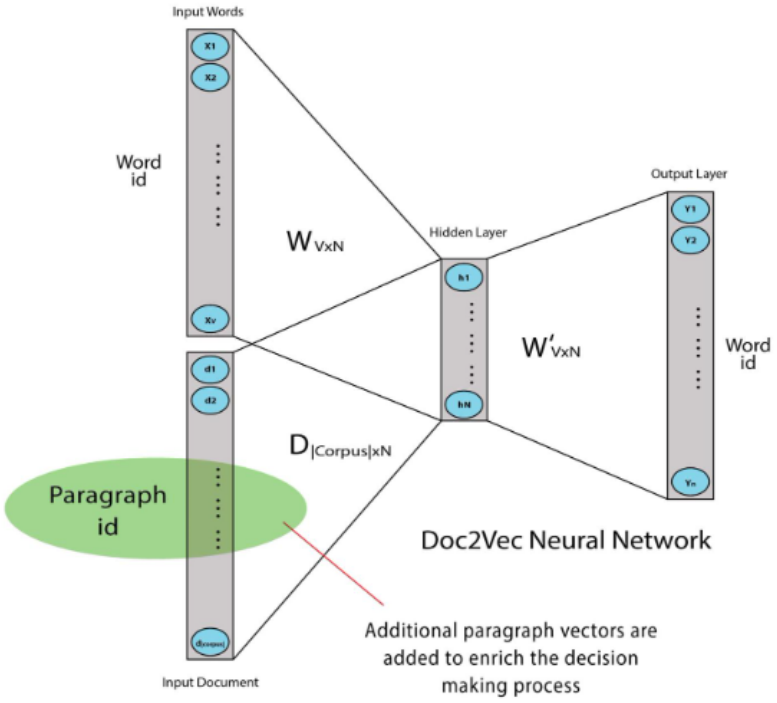

Figure 2. The *Doc2Vec* neural network architecture.

GloVe: The study by Pennington et al. [25] introduced the count-based *GloVe* (Global Vectors) model, which learns vectors through dimensionality reduction on a co-occurrence counts matrix. The large co-occurrence matrix

of words (rows) and context (columns) maps how frequently each word appears in a certain context in a large corpus. The *GloVe* technique, which generates word vector spaces with meaningful substructure, has been found to achieve state-of-the-art performance on several text processing experiments, such as word similarity tasks, word analogy tasks, and Named Entity Recognition (NER).

FastText: *FastText* is a Facebook Artificial Intelligence Research (FAIR) open-source library for NLP [26]. The *FastText* approach combines several robust ML and NLP techniques, including a bag of words and a bag of n-grams, as well as using sub-word information and sharing information across classes through a hidden representation to achieve efficient text classification. *FastText* is on par with some deep learning frameworks in terms of accuracy in learning word vector representations, but it is much faster to train, making it more scalable to larger datasets. Its use of sub-word n-gram processing has been shown to increase *FastText*'s accuracy over *Word2Vec* when handling uncommon words or words that are out-of-vocabulary.

A few cross-comparison studies evaluating the strengths and weaknesses of *Word2Vec*, *FastText*, and *GloVe* have generally concluded that they are overall comparable to each other, although performance varies depending on the task and the length of text in both queries and target documents [38,39]. A limitation of *Word2Vec*, *FastText*, and *GloVe* is that they only encode vector representations for single words, which do not capture the enriched data provided through more complex, multiple-word structures, such as sentences or paragraphs. In the literature, some researchers have focused on the benefits of adapting traditional, single-word embedding models to learn vectors for multiword expressions [40]. In a few studies, researchers have explored new semantic search models based on SentenceBERT.

SentenceBERT (SBERT): It employs Siamese and triplet architecture to learn semantically meaningful sentence embeddings. In a Siamese network, identical subnetworks process each input, ensuring that semantically similar sentences are projected to nearby points in the embedding space [41].

2.1.3. Similarity Measures

A well-established approach to computing similarity between sentences or documents is to evaluate the cosine similarity or inner product of the centroids of word embeddings (generated from techniques such as *Word2Vec* or *GloVe*) [42]. These document similarity measures have been useful for general clustering and classification of overall topics at a document level. However, simple centroid approximation is regarded as insufficient for calculating the distances between queries and target statements or documents [27]. As queries tend to be short compared to the documents they are being compared against, a lossy centroid approach that calculates the average distance between a query and a document will be less accurate than an approach that searches directly for the query words [43]. The centroid approach also struggles with documents that consist of multiple different topics.

Word Mover's Distance (WMD) [27] emerged from a statistical approach known as Earth Mover's Distance (EMD), which has been successfully applied to computer vision tasks such as image comparison. EMD measures the distance between two probability distributions over a region, and, similarly, the constituent word-vectors of sentences or paragraphs can be considered as distributions or 'piles of meaning' around their individual vector coordinates. WMD has been specifically developed to measure the similarity between two bodies of text (sentences/paragraphs) by calculating the minimum 'travelling distance' between text objects (sentences) as a measure of the sum-of-distances (cosine distance) or the effort it takes to move from one word vector pile configuration to another.

2.1.4. Related Work

The successful application of word embeddings to ad-hoc Information Retrieval (IR) tasks remains a key area of research activity. Nalisnick et al. [42] provided a useful example of the benefits of word embeddings when they described the scenario of querying a document with a high occurrence of the term 'automobile' with a query term 'car'. Techniques such as TF-IDF scored the document relatively low since the term 'car' does not feature prevalently in the document, whereas a word embedding approach scored the document higher because the vector representation for 'automobile' and 'car' are close to each other in the embedding space. Such has been the success and impact of neural word embeddings in NLP tasks that they are now recognized as the main driver of the renewed interest and breakout of NLP in the past few years [44]. These neural word embeddings have become the default representations in many text processing pipelines and neural network architectures, serving as the first layer of

pre-processing that converts raw word tokens into more useful representations [35,44–46].

Bonetti [47] proposed a hybrid model using BM25 for keyword-based search and SentenceBERT for semantic search. It was observed that the proposed model outperformed the limitations of previous approaches. SBERT was also used to encode queries into sentence vectors and measure similarity with cosine similarity against stored queries [48]. In Walsh and Andrade [49], the authors fine-tuned SBERT on the NASA Lessons Learned Information System (LLIS) and analyzed that it performed better than a pre-trained baseline. The authors observed that domain-specific tuning enables accurate semantic search. Several recent studies take advantage of the dual process of leveraging the benefits of word embeddings with WMD similarity ranking. For computing similarities between documents, the WMD approach has been reported to yield lower classification errors when used in conjunction with distance-based classifiers [27]. Combining neural word embeddings with a WMD similarity mechanism was also found to outperform a BM25 ranking system on the TREC 2006 and 2007 Genomics benchmark sets [50,51] using solely semantic comparison as the ranking feature [43].

Recent advances in neural representation learning improve similarity search and show how transformer-based embedding methods can generalize beyond the text IR to multimodal retrieval. Tutor-augmented GANs enhance the robustness and relevance of search pipelines, thereby strengthening the similarity-based retrieval of documents and webpages [52]. Lo et al. [53] analyzed how natural language-conditioned graph generation can improve similarity search over small graphs by producing embeddings aligned with structural properties. According to the literature, it has been observed that dense representations and generative representations can improve the similarity search.

2.2. Research Gaps and Research Objectives

Despite strong baselines like TF-IDF/BM25, LSA advancements, and neural embeddings, current methods still fail to capture fine-grained syntactic or compositional cues, especially for short, partial, or paraphrastic queries. While WMD addresses lexical mismatch, its integration with diverse embedding families, such as *Word2Vec*, *FastText, GloVe*, and *Doc2Vec*, remains underexplored. There is a lack of systematic, head-to-head evaluations across exact-match and paraphrase-oriented IR settings. Additionally, there is limited guidance on when global (LSA) vs. local (skip-gram) vs. transport-based measures are preferable under varying query lengths and linguistic divergence. The following objectives have been framed to address these research gaps:

- Conduct a comparative analysis on different semantic representation strategies for query-statement matching.
- Evaluate IR performance on exact syntactic query-statement matching and paraphrase identification under partial or linguistically divergent content matching.
- Establish baselines using the Global Matrix Factorization approach, LSA and Local Context Window (skip-gram) algorithms, such as *Doc2Vec*, *Word2Vec*, *FastText*, and *GloVe*.
- Investigate Word Movers' Distance (WMD) as an alternative similarity metric over neural word embeddings.
- Compute similarity via distances between individual words of queries and statements.
- Provide, to the best of our knowledge, one of the first evaluations combining WMD with different neural word embedding for semantic IR and similarity ranking.

3. Methodology

This section describes the experimental setup, evaluation dataset, pre-processing, and evaluation metrics implemented to assess the accuracy of each semantic-based information retrieval system. The information retrieval performance of four state-of-the-art semantic representation techniques—*Word2Vec*, *Doc2Vec*, *FastTex*t, and *GloVe*—is compared with that of a traditional vector-based LSA model. The word embeddings from *Word2Vec, FastText,* and *GloVe* were processed using a Word Movers' Distance (WMD) document similarity algorithm to assess the effects on statement similarity rankings compared to three *Doc2Vec* models that vary in word order and contextual analysis.

3.1. Dataset

The evaluation dataset was prepared from a publicly available 2013 IPO Prospectus for Foxtons Estate Agents of London [54]. The Prospectus consists of 223 pages of company and financial data, totaling 141,171 words over 8,127 individual statements or paragraphs. Twelve statements were taken from the Prospectus to be used in 12

separate testing trials. For each of the 12 statements, a set of queries was developed that were syntactic variations of the initial statement. Each search trial query began with the original statement taken from the prospectus. The additional query statements are paraphrased variations of the original statement in that they are constructed differently, but they convey similar meaning.

Each retrieval model's sensitivity to statement length was evaluated by including statements of varying lengths, from a single sentence containing 10 words to multiple sentence statements composed of eight sentences and 215 words, for the retrieval tasks. Additionally, some search queries contained only a portion of the original statement to test if the models would return the paragraph within which the statement snippet occurs. Altogether, there were 59 separate search statements across the 12 trials. An example of the 12 statements and their corresponding query variations is presented in **Table 1**.

Table 1. Target statement: Michael Brown is the Chief Executive Officer of the company.

Query	Target Statement
Query 1	Michael Brown is the Chief Executive Officer of the Company
Query 2	The Chief Executive Officer of the Company is Michael Brown
Query 3	Chief Executive Officer of the Company
Query 4	Michael Brown
Query 5	Chief Executive Officer

Note: Query 1 is the original statement, and Query 2 swaps the name of the person with his job title. Query 3 searches for the last section of the statement only, whereas Query 4 is restricted to the name of the CEO. Finally, Query 5 restricts the search to the actual job title.

3.2. Text Pre-Processing

To ensure fair comparison, the same pre-processing steps were applied to all information retrieval models. Firstly, the raw text string was converted to lower-case. The string was subsequently tokenized by splitting it into the sub-unit words, and paragraph returns were treated as delimiters to specify the text boundaries. The final stage of pre-processing involved removing all common English stop words. As several of the models leverage the use of sub-words or sub-grams, it was decided that stemming and lemmatization would not be applied to the text.

3.3. LSA Baseline Model

The document index for the LSA model was created using the IPO prospectus as the training material to generate the word-to-paragraph matrices. The similarity of the query vectors to the vectors in the document space was measured using cosine similarity.

3.4. Embedding Models

Based on experimental recommendations from Kusner et al. [27,55], it was decided to employ robust, pre-trained, general-purpose word embedding models as opposed to corpus or domain-specific frameworks. These models are trained over vast amounts of data, thus providing a wide diversity of contexts for each word during training. This ensured that the models were not oversensitive to the test dataset.

Word2Vec vectors were generated from the Google News dataset (300 dimensions trained on 100 x 10^9 tokens with a vocabulary size of 3×10^6) [56].

GloVe vectors were generated from the Common Crawl dataset (300 dimensions trained on 840 x 10^9 tokens with a vocabulary size of 2.2×10^6) [57].

FastText vectors were generated from the Common Crawl dataset (300 dimensions trained on 840 x 10^9 tokens with a vocabulary size of 2.2×10^6) [58].

Doc2Vec: There are three variations in the *Doc2Vec* experimental setup:

- Paragraph Vectors − Distributed Memory Model (PV − DM)
- Paragraph Vectors − Distributed Bag of Words Model (PV − DBOW)
- Paragraph Vectors + Distributed Bag of Words Model (PV + DBOW)

Each algorithm variation processes text in a different way, placing different emphasis on word order and contextual analysis. These three variations will each be evaluated for strengths and weaknesses.

Paragraph Vectors Distributed Memory Model (PV − DM): These are the original *Doc2Vec* parameters where the additional paragraph vector acts as a distributed memory store of what is missing from the current context or the topic of the paragraph, and functions as an additional pseudo-word ranging over the entire text (sentence, paragraph, or document) participating in all sliding window samples of Word Vectors. In the PV-DM model, the order of words is important, and many contributors believe this to be an advantage over the 'bag-of-words' approach as it preserves more information about the paragraph [24].

Paragraph Vector − Distributed Bag of Words Model (PV-DBOW): Unlike the PV-DM model, the PV-DBOW model adopts a 'bag of words' approach where word order is irrelevant, and no Word Vectors are trained. Instead, Paragraph Vectors are trained to predict words randomly sampled from the paragraph in the output without using local neighboring words. This simplified approach is more efficient as it requires less data storage. It is also a slightly more flexible approach than PV-DM, as word order is not considered important. It is more likely that this model will recognize semantically similar but syntactically different text [24].

Paragraph Vectors + Distributed Bag of Words Model (PV + DBOW): To produce more consistent results across multiple tasks, Le and Mikolov [24] recommended generating paragraph vectors that are a combination of two vectors: one learned by the standard PV-DM model, and one learned by the PV-DBOW approach. This combinatorial approach of the previous two methods implements simultaneous training of both Paragraph Vectors over the whole text and skip-gram Word Vectors (bag-of-words) over each sliding context window. This approach is slower as the additional training is computationally expensive; however, the benefits of placing both Word Vectors and Paragraph Vectors into the same space enhance the expressiveness and interpretability of the Paragraph Vectors due to their closeness to words of known meanings [24].

For the *Doc2Vec* models, the word embeddings and paragraph centroids are calculated for the prospectus training set. The centroid is then computed for each query, and the statements or paragraphs with the top 20 nearest centroids (in terms of cosine similarity) to the query are retrieved.

3.5. WMD and Similarity Ranking

As an alternative approach to measuring similarity with query-statement centroids of word embeddings, this research evaluates WMD [27] as a means of evaluating similarity through the distance between individual words of queries and statements. WMD uses pre-trained embeddings to compute distance. Let y_i be the embedding of the word *i*. WMD defines the distance between the word *i* and word *j* as $d(i,j) = \left\|y_i - y_j\right\|_2$, also known as transport cost from word *i* to word *j*. WMD also assigns a flow f_i to each word *i* which can be defined as given in Equation (1).

$$f_i = \frac{n(i)}{|f|},\ |f| = \sum_i n(i) \tag{1}$$

where $|f|$ is the total word count of a text sequence. Then, WMD measures the dissimilarity of two texts by computing the minimum total cost to move all words mass from one text to another by using the following Equation (2).

$$\min_{P_{i,j} \geq 0} \sum_{i \in I} \sum_{j \in J} P_{i,j} d(i,j) \tag{2}$$

Subject to:

$\sum_{j \in J} P_{i,j} = f_i \ \forall i \in I,$ and $\sum_{i \in I} P_{i,j} = f'_j \ \forall j \in J,$

where *I* and *J* are set of words in text sequences d_1 and d_2 respectively. $P_{i,j}$ represents the amount of flow that travels from word *i* to word *j.*

It was decided to test the WMD model across several well-established pre-trained vector libraries to determine the best combination for effective semantic information retrieval. The three model combinations were as follows:

- WMD + *Word2Vec*
- WMD + *FastText*
- WMD + *GloVe*

The text from the IPO Prospectus and the query text used to search the document were encoded as vectors through the above word embedding techniques before applying query-statement similarity comparisons with the WMD architecture. Applying the different similarity comparisons across all techniques (LSA, *Doc2Vec*, and WMD variations), a set of relevance scores is generated for each query-statement pair. The relevancy scores are ranked, and consideration is limited to the top 20 ranked statements.

4. Results and Evaluation

To evaluate the robustness and accuracy of each semantic retrieval method, three separate comparisons were made regarding the ranking of correct statement matches (**Table 2**). The first comparison assessed the number of correct matches that each model returned within the top 20 ranked positions when similarity ranking was applied. The second comparison assesses the number of correct matches that rank within the top three positions when similarity metrics are applied. Finally, the last evaluation assesses how many target statements each model returns to the top-ranked position when similarity ranking is applied.

Table 2. The number and percentage of correct statement matches (top 20 similarity ranked positions).

System	# Correct Statement Matches	% Correct Statement Matches
WMD-*GloVe*	59	100%
WMD-*FastText*	59	100%
WMD-*Word2Vec*	59	100%
PV + DBOW	53	89.83%
PV-DBOW	40	67.8%
PV-DM	33	55.93%
LSA	11	18.64%

Note: The number and percentage of correct statement matches returned within the top 20 similarity ranked positions by each semantic retrieval model across 12 query-statement trials totaling 59 comparison statements.

The WMD-*GloVe*, WMD-*FastText*, and WMD-*Word2Vec* systems outperform the other text comparison systems, returning 100% correct matches within the top 20 ranked results, as shown in **Table 1**. The only system that comes close to this performance is the *Doc2Vec* PV + DBOW method, which returns 89.83% (53/59) correct matches. Upon further analysis, it was discovered that all three systems return 100% correct statement matches within the top 10 results. Amongst the *Doc2Vec* models, the PV-DBOW version of *Doc2Vec* outperforms the original PV_DM model, returning 40/59 (67.8% return rate) statement matches compared to 33/59 (55.93% return rate). However, the PV + DBOW *Doc2Vec* model is the clear winner, with a return rate of 89.83%, returning 53 correct statement matches out of a possible 59. This stronger performance of the PV + DBOW *Doc2Vec* model aligns with the findings and recommendations of Le and Mikolov [24]. The combination of both low-level skip-gram Word Vectors with higher-level Paragraph Vectors harnesses greater semantic expressiveness when they work together in the same distribution space.

The superior performance of the 'bag-of-words' *Doc2Vec* PV-DBOW approach compared to the PV-DM model indicates that it is much more flexible to the change of word ordering that occurs in many of the trials compared to the more rigid PV-DM sliding paragraph window method. In this case, the benefit of preserving additional information about the paragraph through the sliding window technique is outweighed by the 'bag-of-word' approach, where word order is not important. This failure to generalize rephrased paragraphs indicates that PV-DM may be too restrictive when attempting to match statements that are semantically similar yet textually different.

The LSA model achieves a very poor return of 11 correct statement matches (18.64%) in the top 20 relevancy-ranked positions. Comparing this baseline model to the more advanced word and paragraph embedding approaches can be considered in broader terms as a comparison between two main approaches to learning and generating word vectors. LSA is a Global Matrix Factorization method that processes text at the higher document level compared to the Local Context Window (skip-gram) approaches, such as *Doc2Vev* variations and word embedding algorithms. The poor performance of the LSA model reinforces evidence that global approaches fail to efficiently leverage the lower-level statistical information that exposes the links between component vectors. Global techniques tend to work better at the document level processing, such as topic clustering or classification. Indeed, from informal qualitative analysis of the top 20 results returned for each LSA query, there is no obvious semantic consistency or relatedness in the content or themes of the returned statements. All WMD systems have a 100% record for returning

top 20 ranked statement matches.

However, when analyzing performance in terms of the number of top three ranked returned statement hits, a subtle difference in the accuracy performance of WMD-*GloVe* and WMD-*FastText* compared to WMD-*Word2Vec* emerges (**Table 3**). The WMD-*GloVe* system outperforms all other systems in the top 3 ranking comparisons by returning an impressive 100% (59/59) top 3 ranked search results for all possible search term combinations compared to 58/59 (98.31%) for the WMD-*FastText* system and 56/59 (94.92%) for the WMD-*Word2Vec* system. The robust ranking performance of all three WMD systems is contrasted against a significant drop in performance of the *Doc2Vec* models when the percentage of top 3 ranking returned statements is considered. The poor performance of the *Doc2Vec* architectures on shorter query-statement combinations aligns with evidence from other research indicating that paragraph embedding approaches are best suited to longer text segments [59]. The developers of *Doc2Vec* [24] note that very short texts tend not to generate useful representations from this model. If performance on shorter paragraphs or sentences is important, they suggest factoring in some mechanism to outweigh them. The authors propose a method of repeating a paragraph that is 1/Nth the average size by N times randomly throughout the training set or implementing N times more steps during inference. The mediocre performance of the LSA system is further compounded by the lowest return of 5 statement matches (8.47%) in the top 3 ranked returned results. All WMD systems recorded strong performances with the top 30 and top three ranked comparisons.

Table 3. The number and percentage of correct statement matches (top 3 similarity ranking positions).

System	# Correct Statement Matches	% Correct Statement Matches
WMD-*GloVe*	59	100%
WMD-*FastText*	58	98.31%
WMD-*Word2Vec*	56	94.92%
PV + DBOW	47	79.66%
PV-DM	25	42.37%
PV-DBOW	23	38.98%
LSA	5	8.47%

Note: The number and percentage of correct statement matches returned within the top 3 similarity ranking positions by each semantic retrieval model across 12 query-statement trials totaling 59 comparison statements.

However, when analyzing performance in terms of the number of top-ranked (number one) returned statement hits, a significant difference in the accuracy performance of WMD-*GloVe* and WMD-*FastText* compared to WMD-W*ord2Vec* becomes apparent (**Table 4**). The WMD-*GloVe* system returns an impressive 53/59 (89.83%) top-ranked statement matches compared to a 50/59 (84.74%) return rate for the WMD-*FastText* system and a significantly lower 18/59 (58.98%) return rate for the WMD-*Word2Vec* system. In fact, the WMD-*Word2Vec* system is outperformed in these trials by the *Doc2Vec* PV + DBOW system, which returns 39/59 (66.10%) top-ranked statement matches. On further analysis, the WMD-*GloVe* system was also found to return all search matches (100%) in the top 2 ranked positions compared to 56/59 (94.92%) top 2 ranked returns for WMD-*FastText*, 51/59 (86.44%) top 2 ranked returns for WMD-Google, and 44/59 (74.58%) top 2 ranked returns for the *Doc2Vec* PV + DBOW system.

Table 4. The number and percentage of correct statement matches (number one ranked similarity ranking positions).

System	# Correct Statement Matches	% Correct Statement Matches
WMD-GloVe	53	89.83%
WMD-FastText	50	84.74%
PV + DBOW	39	66.10%
WMD-Word2Vec	18	58.98%
PV-DBOW	16	27.12%
PV-DM	15	25.42%
LSA	2	0.03%

Note: The number and percentage of correct statement matches returned within the number one ranked similarity ranking positions by each semantic retrieval model across 12 query-statement trials totaling 59 comparison statements.

From the ranked similarity results as shown in **Figure 3**, it has been established that WMD-*GloVe* achieves the most robust accuracy performance for statement matching from re-worded and partial queries. The real power of this vector-based approach, however, lies in its ability to achieve semantic matching of statements closely related to the theme or content. The superior performance of WMD + *GloVe* can be attributed to a synergy between *GloVe*'s

globally consistent embedding space and WMD's reliance on stable geometric structure. Unlike *Word2Vec*, which derives embeddings from local context windows and can produce higher variance representations [45,52], *GloVe* constructs word vectors by factorizing a global co-occurrence matrix across the entire corpus [25]. This global statistical grounding yields a more consistent semantic space, which aligns well with WMD's optimal transport mechanism for measuring document distance [27]. As a result, *GloVe* + WMD benefits from both robust semantic stability and precise word-level alignment, enabling more accurate document comparisons. The key assumption is that *GloVe*'s aggregated global context captures semantic regularities more reliably than locally predictive methods, thereby reducing noise and enhancing WMD's effectiveness, particularly in top-ranked retrieval scenarios.

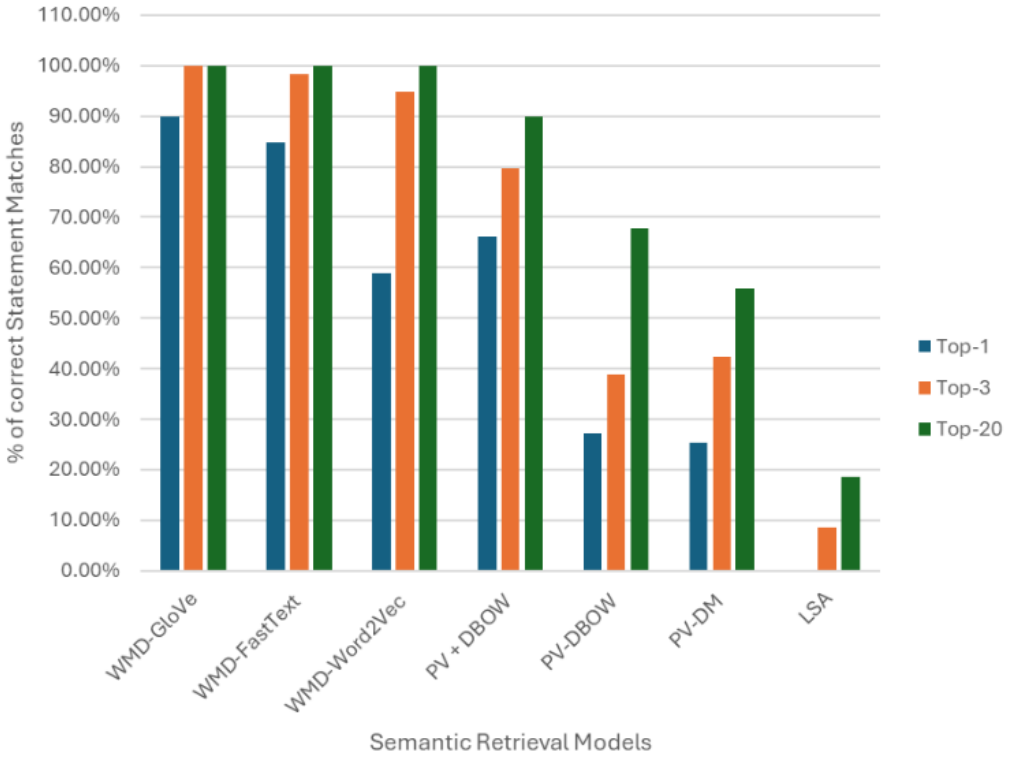

Figure 3. Comparison of percentage of correct statement matches returned by each model.

To highlight this semantic processing ability, an informal qualitative analysis of the trial 2 results was conducted to assess the relatedness of the top 20 statements returned based on the following query: "Michael Brown is the Chief Executive Officer of the Company". This query search was analyzed as the WMD-based systems were the only models that achieved 100% recall on it. Using the search phrase, 10 out of the top 20 ranked statements (including the top 4 ranked statements) specifically mention "Michael Brown" as the CEO or executive director of the company. While this indicates successful term matching at a syntactic level, the contents of the remaining 10 statements demonstrate text matching at a semantic level by the combined WMD and *GloVe* word embeddings model. The remaining 10 statements reference semantically similar topics or concepts, including management structure, key management personnel, board of directors, and other details relating to company management. This ability to cluster semantically related concepts within the distribution space has important implications for facilitating users' search experiences through informed query expansion and providing relevant responses to partial or incomplete query searches.

5. Discussion

Our results show that pre-trained word embedding models can be effectively applied to different domains with considerable success. However, there are certain use cases and business domains where the language used is

particularly nuanced and specific. In these circumstances, being uncoupled from the domain ontologies would be a disadvantage, leading to systems that fail to understand the information needs of the end-users. It is therefore necessary to adopt embedding schemes that are enriched by exploiting domain-tailored knowledge. The benefit of the unsupervised learning algorithms evaluated in this paper is that they can be trained on domain-specific ontology and lexical resources without the need for time-consuming supervised learning prerequisites, such as term extractors or manually labelled training data. The power of these algorithms is that their term x object matrices can be automatically populated for any text collection where the underlying concepts can be identified by completely automatic statistical processes. For example, in these experimental trials, pre-trained *GloVe* vectors were used. They proved successful in the accurate retrieval of both short and long query-statement searches. If, however, domain-specific tuning was deemed necessary for specialized datasets, the *GloVe* model is much more efficient to train compared to the *Doc2Vec* variations. Scalability is facilitated as it is easier to parallelize the implementation, enabling it to train over more data, and populating the co-occurrence matrix requires a single pass through an entire corpus to collect the statistics.

The accuracy and robustness of these vector-based semantic retrieval models have set the agenda for this analysis paper. However, for practical applications, these mechanisms also need to be considered within a wider context of a semantic search framework and a semantic support infrastructure where usability and the support of the end-user are the focus of industrial information retrieval and management solutions. Given that a substantial performance gap remains between Information Retrieval systems and what users need and expect from them [9] and considering the volume of time that users can spend on searching tasks as part of their everyday work duties, it is necessary to consider an entire search ecosystem built around semantic search. Many large businesses have complex information spaces that require additional support for the end user to facilitate their search strategies for navigating these unfamiliar underlying knowledge structures. This is particularly necessary if they are seeking information from a wide range of topics or have uncertainty about the nature of the search problem. This lack of definition and certainty about the keywords to choose further highlights the need to factor semantic synonymy into the equation. Semantic search will be at the core of this new breed of techniques that are being developed to support the browsing of information spaces. Apart from text-based semantic search, which has been at the center of this investigation, Information Retrieval accuracy can be further improved when we incorporate the semantic-based approaches into a hybrid framework of search strategies.

Other measures of algorithmic relevance can be included in the overall suite of search tools, including network statistics, click-through data, and semantic-driven query expansion [9]. Query expansion is an iterative and exploratory process where the user actively engages with the search system to refine their queries in response to the results that are returned. This can be seen on the interfaces of many web search engines where expanded query suggestions appear in response to users' input. The use of query expansion has been found to increase recall, and this process can be enriched through semantic-based query recommendations or auto-suggest. Here, the query expansion system would use the word embedding techniques to suggest potential query terms based on semantic similarity or synonymous names for concepts [33,60,61].

6. Conclusions

This investigation demonstrates the ability of semantic or conceptual-based search strategies to exploit the latent underlying semantic structure of text and how this can be leveraged to improve the quality and relevancy of the search experience. Semantic search enables the retrieval of documents based on how similar the concepts in the query are to the concepts in the document. These concepts represent high-level ideas in each domain. Semantic representation strategies can be viewed as a means of narrowing the gap between the mismatch of words that are contained in documents and the words expressed in queries, reflecting users' intentions. For the modern user, this intuitive behavior of semantic search has almost become expected, thanks to services such as Google, as searchers expect search engines to 'do as I mean – not as I say' when they query. Semantic search enables users to get relevant results even when they input shorthand, truncated, or misspelled queries containing only a few keywords. Overcoming lexical problems, such as misspellings and partial queries, facilitates data exploration and enables users to find target text in large collections of data, allowing them to interact more fully with the data and enrich the information discovery process.

Limitations and Future Work

- This study is limited to a single domain-specific dataset (the Foxtons IPO prospectus), and our findings, therefore, can be interpreted as a focused case study in corporate finance documents. While this study provides robust proof-of-concept, the methodology's broader applicability requires validation on contrasting domains. Future work will extend this evaluation to such heterogeneous datasets to assess the generalizability of our approach.
- One of the limitations of this work is the absence of benchmarking against Transformer-based models, such as SBERT and GAN-VAE developments [49,52,53]. We frame this study as a focused case study on static embeddings with WMD and identify SBERT and GAN-VAE comparison as an important direction for future research.

Author Contributions

Conceptualization, E.O.C., N.M., S.R., K.C., E.M., A.C., and A.B.; methodology, E.O.C.; software, E.O.C., N.M., and S.R.; validation, E.O.C., N.M., and S.R.; formal analysis, E.O.C., N.M., S.R.; investigation, E.O.C.; resources, E.O.C.; data curation, E.O.C., N.M., and S.R.; writing—original draft preparation, E.O.C.; writing—review and editing, N.M., S.R., K.C., E.M., A.C., and A.B.; visualization, E.O.C.; supervision, K.C., E.M., A.C., and A.B.; project administration, K.C., E.M., A.C., and A.B. All authors have read and agreed to the published version of the manuscript.

Funding

This work received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

No new data were generated or analyzed in this study. All data used are from publicly available sources cited within the manuscript.

Acknowledgments

This research was supported and co-authored by Allen Overy Shearman Sterling LLP as part of the continuing work of the Ulster University Legal Innovation Centre. This Centre provides research, development, and educational resources for the promotion of innovation in legal services provision and access to justice.

Conflicts of Interest

The authors declare that there is no conflict of interest.

References

1. Cakir, A.; Gurkan, M. Modified Query Expansion Through Generative Adversarial Networks for Information Extraction in E-Commerce. *Mach. Learn. Appl.* **2023**, *14*, 100509.
2. Růžička, M.; Novotný, V.; Sojka, P.; et al. Flexible Similarity Search of Semantic Vectors Using Fulltext Search Engines. In Joint Proceedings of the International Workshops on Hybrid Statistical Semantic Understanding and Emerging Semantics, and Semantic Statistics (Hybrid-SemStats), Vienna, Austria, 22 October 2017.
3. Cohen, W.W. Data Integration Using Similarity Joins and a Word-Based Information Representation Language. *ACM Trans. Inf. Syst.* **2000**, *18*, 288–321.
4. Schallehn, E.; Sattler, K.U.; Saake, G. Efficient Similarity-Based Operations for Data Integration. *Data Knowl. Eng.* **2004**, *48*, 361–387.

5. Madhavan, J.; Bernstein, P.A.; Doan, A.; et al. Corpus-Based Schema Matching. In Proceedings of the 21st International Conference on Data Engineering (ICDE'05), Tokyo, Japan, 5–8 April 2005.
6. Gries, S.T. Polysemy. Chapter 2: Polysemy. In *Cognitive Linguistics: Key Topics*; Dabrowska, E., Divjak, D., Eds.; De Gruyter Mouton: Berlin, Germany, 2019; pp. 23–43.
7. Al-Smadi, M.; Jaradat, Z.; Al-Ayyoub, M.; et al. Paraphrase Identification and Semantic Text Similarity Analysis in Arabic News Tweets Using Lexical, Syntactic, and Semantic Features. *Inf. Process. Manag.* **2017**, *53*, 640–652.
8. Brokos, G.I.; Malakasiotis, P.; Androutsopoulos, I. Using Centroids of Word Embeddings and Word Mover's Distance for Biomedical Document Retrieval in Question Answering. In Proceedings of the 15th Workshop on Biomedical Natural Language Processing, Berlin, Germany, 12 August 2016.
9. van Opijnen, M.; Santos, C. On the Concept of Relevance in Legal Information Retrieval. *Artif. Intell. Law* **2017**, *25*, 65–87.
10. Maxwell, K.T.; Schafer, B. Concept and Context in Legal Information Retrieval. In *Legal Knowledge and Information Systems*; Francesconi, E., Sartor, G., Tiscornia, D., Eds.; IOS Press BV: Amsterdam, Netherlands, 2008; pp. 63–72.
11. Mikolov, T.; Le, Q.V.; Sutskever, I. Exploiting Similarities Among Languages for Machine Translation. *arXiv preprint* **2013**, *arXiv:1309.4168*.
12. Dumais, S.T. Latent Semantic Analysis. *Annu. Rev. Inf. Sci. Technol.* **2005**, *38*, 188–230.
13. Vani, K.; Gupta, D. Unmasking Text Plagiarism Using Syntactic-Semantic Based Natural Language Processing Techniques: Comparisons, Analysis and Challenges. *Inf. Process. Manag.* **2018**, *54*, 408–432.
14. Zhang, C.; Zhang, L.; Wang, C.J.; et al. Text Summarization Based on Sentence Selection With Semantic Representation. In Proceedings of the 26th International Conference on Tools with Artificial Intelligence, Limassol, Cyprus, 10–12 November 2014.
15. Erkan, G.; Radev, D.R. LexRank: Graph-Based Lexical Centrality as Salience in Text Summarization. *J. Artif. Intell. Res.* **2004**, *22*, 457–479.
16. Lapata, M.; Barzilay, R. Automatic Evaluation of Text Coherence: Models and Representations. In Proceedings of the International Joint Conference on Artificial Intelligence, Edinburgh, UK, 30 July–5 August 2005.
17. Wegrzyn-Wolska, K.; Szczepaniak, P.S. Classification of RSS-Formatted Documents Using Full Text Similarity Measures. In Proceedings of the 5th International Conference, ICWE 2005, Sydney, Australia, 27–29 July 2005.
18. Gella, S.; Keller, F.; Lapata, M. Disambiguating Visual Verbs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *41*, 311–322.
19. Liu, T.; Guo, J. Text Similarity Computing Based on Standard Deviation. In Proceedings of the International Conference on Intelligent Computing, ICIC 2005, Hefei, China, 23–26 August 2005.
20. Jin, P.; Zhang, Y.; Chen, X.; et al. Bag-of-Embeddings for Text Classification. In Proceedings of the 25th International Joint Conference on Artificial Intelligence, New York, NY, USA, 9–15 July 2016.
21. Darraz, N.; Karabila, I.; El-Ansari, A.; et al. Integrated Sentiment Analysis with BERT for Enhanced Hybrid Recommendation Systems. *Expert Syst. Appl.* **2025**, *261*, 125533.
22. Grainger, T.; Potter, T.T. *Solr in Action*. Manning Publications: Shelter Island, NY, USA, 2014.
23. Mikolov, T.; Sutskever, I.; Chen, K.; et al. Distributed Representations of Words and Phrases and Their Compositionality. In Proceedings of the Advances in Neural Information Processing Systems 26 (NIPS 2013), Lake Tahoe, NV, USA, 5–11 December 2013.
24. Le, Q.; Mikolov, T. Distributed Representations of Sentences and Documents. In Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 21–26 June 2014.
25. Pennington, J.; Socher, R.; Manning, C.D. 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.
26. Bojanowski, P.; Grave, E.; Joulin, A.; et al. Enriching Word Vectors With Subword Information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146.
27. Kusner, M.; Sun, Y.; Kolkin, N.; et al. From Word Embeddings to Document Distances. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015.
28. Tinega, G.A.; Mwangi, W.; Rimiru, R. Text Mining in Digital Libraries Using OKAPI BM25 Model. *Int. J. Comput. Appl. Technol. Res.* **2019**, *7*, 398–406.
29. Momtazi, S. Unsupervised Latent Dirichlet Allocation for Supervised Question Classification. *Inf. Process. Manag.* **2018**, *54*, 380–393.

30. Evangelopoulos, N.E. Latent Semantic Analysis. *Wiley Interdiscip. Rev. Cogn. Sci.* **2013**, *4*, 683–692.
31. Hyung, Z.; Park, J.S.; Lee, K. Utilizing Context-Relevant Keywords Extracted From a Large Collection of User-Generated Documents for Music Discovery. *Inf. Process. Manag.* **2017**, *53*, 1185–1200.
32. Islam, A.; Inkpen, D. Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity. *ACM Trans. Knowl. Discov. Data* **2008**, *2*, 1–25.
33. Uren, V.; Lei, Y.; Lopez, V.; et al. The Usability of Semantic Search Tools: A Review. *Knowl. Eng. Rev.* **2007**, *22*, 361–377.
34. Bruni, E.; Tran, N.K.; Baroni, M. Multimodal Distributional Semantics. *J. Artif. Intell. Res.* **2014**, *49*, 1–47.
35. Baroni, M.; Dinu, G.; Kruszewski, G. Don't Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 22–27 June 2014.
36. Bengio, Y.; Ducharme, R.; Vincent, P.; et al. A Neural Probabilistic Language Model. *J. Mach. Learn. Res.* **2003**, *3*, 1137–1155.
37. Schnabel, T.; Labutov, I.; Mimno, D.; et al. Evaluation Methods for Unsupervised Word Embeddings. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015.
38. Muneeb, T.H.; Sahu, S.; Anand, A. Evaluating Distributed Word Representations for Capturing Semantics of Biomedical Concepts. In Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015), Beijing, China, 30 July 2015.
39. Cao, S.; Lu, W. Improving Word Embeddings With Convolutional Feature Learning and Subword Information. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
40. Bagheri, E.; Ensan, F.; Al-Obeidat, F. Neural Word and Entity Embeddings for Ad Hoc Retrieval. *Inf. Process. Manag.* **2018**, *54*, 657–673.
41. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019.
42. Nalisnick, E.; Mitra, B.; Craswell, N.; et al. Improving Document Ranking with Dual Word Embeddings. In WWW'16: 25th International World Wide Web Conference, Montreal, Canada, 11–15 Apr 2016.
43. Kim, S.; Fiorini, N.; Wilbur, W.J.; et al. Bridging the Gap: Incorporating a Semantic Similarity Measure for Effectively Mapping PubMed Queries to Documents. *J. Biomed. Inform.* **2017**, *75*, 122–127.
44. Goth, G. Deep or Shallow, NLP Is Breaking Out. *Commun. ACM* **2016**, *59*, 13–16.
45. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828.
46. Levy, O.; Goldberg, Y. Neural Word Embedding as Implicit Matrix Factorization. In Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NeurIPS 2014), Montreal, Canada, 8–13 December 2014.
47. Bonetti, L. Design and Implementation of a Real-World Search Engine Based on Okapi BM25 and Sentence-BERT. Master Thesis, University of Bologna, Bologna, Italy, 2021.
48. Sharma, K.V.; Ayiluri, P.R.; Betala, R.; et al. Enhancing Query Relevance: Leveraging SBERT and Cosine Similarity for Optimal Information Retrieval. *Int. J. Speech Technol.* **2024**, *27*, 753–763.
49. Walsh, H.S.; Andrade, S.R. Semantic Search with Sentence-BERT for Design Information Retrieval. In Proceedings of the International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE), St. Louis, MO, USA, 14–17 August 2022.
50. Hersh, W.R.; Cohen, A.M.; Roberts, P.M.; et al. TREC 2006 Genomics Track Overview. In Proceedings of the 15th Text REtrieval Conference (TREC 2006), Gaithersburg, MD, USA, 14–17 November 2007.
51. Cohen, A.; Ruslen, L.; Roberts, P. TREC 2007 Genomics Track Overview. In Proceedings of the 16th Text Retrieval Conference (TREC 2007), Gaithersburg, MD, USA, 5–9 November 2007.
52. Lin, Y.; Ying, C.; Xu, B.; et al. Dual Cycle Generative Adversarial Networks for Web Search. *Appl. Soft Comput.* **2024**, *153*, 111293.
53. Lo, R.; Datar, A.; Sridhar, A. LIC-GAN: Language Information Conditioned Graph Generative GAN Model. *arXiv preprint* **2023**, *arXiv:2306.01937*.
54. Foxtons. Available from: https://ftalphaville-cdn.ft.com/wp-content/uploads/2013/09/Foxtons.pdf (accessed 15 August 2025).
55. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv*

preprint **2013**, *arXiv:1301.3781*.
56. Google. Available from: https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?resourcekey=0-wjGZdNAUop6WykTtMip30g (accessed 15 August 2025).
57. Common Crawl. Available from: https://commoncrawl.org/ (accessed 15 August 2025).
58. Facebook Open Source. Available from: https://fasttext.cc/docs/en/english-vectors.html (accessed 15 August 2025).
59. Galke, A.; Saleh, L.; Scherp, A. Evaluating the Impact of Word Embeddings on Similarity Scoring in Practical Information Retrieval. In *Informatik*; Eibl, M., Gaedke, M., Eds.; Gesellschaft für Informatik: Bonn, Germany, 2017; pp. 2155–2167.
60. Kuzi, S.; Shtok, A.; Kurland, O. Query Expansion Using Word Embeddings. In Proceedings of the CIKM'16: ACM Conference on Information and Knowledge Management, Indianapolis, IN, USA, 24–28 October 2016.
61. Miller, S.; Curran, K.; Lunney, T. Detection of Anonymising Proxies Using Machine Learning. *Int. J. Digit. Crime Forensics* **2021**, *13*, 1–17.