CHAPTER 2
FOUNDATIONS OF SEMANTIC SIMILARITY

This chapter discusses the fundamentals of semantic similarity and classifies the similarity metrics, according to characteristics. It presents the details about the sources involved in measuring the semantic similarity. It also reviews how the semantic similarity involves in information retrieval models. Finally, it explores various semantic similarity approaches utilized in different applications.

2.1 Basis of Similarity

The similarity is an important and widely used concept in NLP, AI, IR and many other fields [3] [14]. It defines the quantity that denotes the strength of the relationship between two objects. The value of a quantity is in the range of either -1 to +1 or 0 to 1. Humans can easily decide whether the pair of words is related or not. For example, most would agree that apple and orange are more similar than Apple and toothbrush. The similarity is the measure used to measure the similarity between objects. The object can be a word, sentences, documents, and ontologies. Semantic similarity measures the similarity between similar entities like apple and orange [8]. These entities are hyponyms of fruit. However, semantic similarity can be applied to measure the similarity between dissimilar entities. For instance, glass and water, tree and shade, or gym and weights are not similar entities but are related by some relationship. This similarity can be measured in several ways using any of the similarity metrics. The following subsections give an overview of the metrics available to measure the similarity and different types of similarity. Diverse measures of similarity are helpful in various types of analysis.

2.1.1 Metrics

Semantic similarity, Semantic distance, and Relatedness are metrics used to measure the similarity. Semantic Similarity Measures concentrate on the resemblance of the two given concepts. Relatedness estimates the relation between the concepts. For instance, “Apple” and “Computer” are related, but not much
similar, but “Apple” and “Fruit” are similar in some degree. Relatedness Measures the more general concepts than the similarity [29]. The semantic distance is the inverse measurement of semantic similarity, and it finds the similarity among the concepts using the knowledge source. The concepts in a knowledge source are organized in a hierarchical structure. The shortest path is the most commonly used measure of semantic distance.

2.1.2 Types

Similarity can be classified into two kinds: Attributional and Relational similarity [15] [29]. The attributional similarity is the measure of similarity between the attributes of two objects. If two objects have similar characteristics, they are considered as the related objects. The relational similarity is the measure that estimates the similarity between the objects, according to its relations. It defines the amount of attributional similarity between two words X and Y, depending on the degree of correspondence between the properties of X and Y. If the objects have a high level of attributional similarity, they are known as synonyms. If the objects have a high degree of relational similarity, they are known as analogous.

2.1.2.1 Relational Similarity

Relational similarity measures the implicit relationship presents between the two pair of objects. For instance, the word pairs (ostrich, bird) and (lion, cat). The ostrich is a large bird and lion is a large cat. The implicitly stated relationship between the word-pair is: “is a large” [15]. The capability to identify the implicitly stated analogies of relational similarity is more helpful in information retrieval to search the required information. However, the process of measuring the implicitly denoted relations among the objects is still a challenging task due to the following reasons: First the relational similarity is not constant, it varies over the time. For instance, initially, two companies can be competitors. Now the relationship is (X competitive Y). After some year, one company acquires the other company. Now the relationship is (X subset Y). In another instance, the word pair can have several relationships. For example, the word pair Ostrich and Bird have a relationship “is a”
and also a “flightless” state. Hence, the relational similarity measure should extract all the possible relations between the word pair to compute the similarity. Third, the relation between the words can be expressed in several ways. The final problem with the word pair can be a named entity like company names and people names. Despite these challenges, the relational similarity measurement is adaptable to NLP applications such as word analogies generation [30] and noun-modifier compound classification.

2.1.2.2 Attributional Similarity

Attributional similarity denotes the relationship between the attributes of the two objects [29]. For example, the concept car and automobile are attributional similarities. Since, they have similar attributes such as doors, four wheels and both are used for transportation. Thus, the concept car and automobile are considered as synonyms. Semantic similarity is a confidence score that reflects the semantic relation between the two objects. Compared to relational similarity, attributional similarity measure has more impact on NLP and IR applications. The primary focus of the rest of the work is to discuss the approaches and metrics used in attributional similarity, and it employs the word semantic similarity technique instead of attributional similarity.

2.2 Semantic Similarity Measures

Semantic similarity measures are tools that calculate quantitatively or qualitatively the degree of similarity between the terms/concepts based on any of the knowledge sources. The knowledge source can be a dictionary, corpus, thesaurus, and ontology. The following are the different types of semantic similarity measures:

2.2.1 Distance-Based Similarity Measures

Distance based similarity measures measure the degree of dissimilarity (distance) between two given objects. The minimum value of dissimilarity is zero. The similarity measure between the objects should be calculated such that the similar objects are grouped together. From the distance measure, the value of
similarity can be calculated using the formula of similarity measure = (1 - similarity
distance). Following are some of the methods to calculate the distance similarity
useful for the information retrieval system [31] [32].

**Minkowski Distance**

Minkowski is the general method to estimate the metric distance for
multidimensional data. The norm Minkowski distance measure is defined as the
distance $D_{ij}$ between two parts $i$ and $j$ as,

$$
D_{ij} = \sqrt[p]{\sum_{i=1}^{d} |P_i - Q_i|^p}
$$

(2.1)

**Manhattan/City Block Distance**

Manhattan/City block distance measures the accurate distance between two
points. It is defined as the Minkowski distance as a norm value of 1.

$$
D_{ij} = \sum_{i=1}^{d} |P_i - Q_i|
$$

(2.2)

**Euclidean Distance**

Euclidean distance is the distance of Minkowski distance at a norm value of 2. It
is the widely used similarity measure to compute the distance between two objects.

$$
D_{ij} = \sqrt{\sum_{i=1}^{d} (P_i - Q_i)^2}
$$

(2.3)

**Chebyshev Distance**

At $n \rightarrow \infty$ Minkowski distance is defined as Chebyshev distance. This measure
indicates the highest distance between two vectors of any coordinate dimension.

$$
D_{ij} = \max_{i} |P_i - Q_i|
$$

(2.4)
Jaccard Distance

Jaccard distance is a simple measure, measures the distance between two sets. It defines the relation of the size of the intersection of the sets and the size of the union of the sets. It is a complementary method for Jaccard coefficient and the value of the Jaccard distance is computed by differentiating the Jaccard coefficient (JC) from 1. Where, P and Q indicate the two sample sets.

\[
S_{Jac} = \frac{\sum_{i=1}^{d} p_i q_i}{\sum_{i=1}^{d} p_i^2 + \sum_{i=1}^{d} q_i^2 - \sum_{i=1}^{d} p_i q_i} \quad \ldots \ldots \ (2.5)
\]

\[
D_{Jac} = 1 - S_{Jac} = \frac{\sum_{i=1}^{d} (p_i - q_i)^2}{\sum_{i=1}^{d} p_i^2 + \sum_{i=1}^{d} q_i^2 - \sum_{i=1}^{d} p_i q_i} \quad \ldots \ldots \ (2.6)
\]

Dice’s Coefficient

Dice’s Coefficient is similar to Jaccard distance. It is twice the number of standard terms between the given two terms (P, Q) and divide by the total number of terms in the given terms to measure the similarity.

\[
S_{Dice} = \frac{2 \sum_{i=1}^{d} p_i q_i}{\sum_{i=1}^{d} p_i^2 + \sum_{i=1}^{d} q_i^2} \quad \ldots \ldots \ (2.7)
\]

\[
D_{Dice} = 1 - S_{Dice} = \frac{\sum_{i=1}^{d} (p_i - q_i)^2}{\sum_{i=1}^{d} p_i^2 + \sum_{i=1}^{d} q_i^2} \quad \ldots \ldots \ (2.8)
\]

Cosine Similarity

Cosine similarity converts the terms (A, B) into a vector; it applies the Euclidean cosine rule to calculate the similarity. It is most commonly used vector-based similarity measure in text mining and information retrieval. This method is employed with other methods to overcome the dimensionality limitation in vector base.
\[
S_{\cos} = \frac{\sum_{i=1}^{d} P_i Q_i}{\sqrt{\sum_{i=1}^{d} P_i^2 \sqrt{\sum_{i=1}^{d} Q_i^2}}} \quad \text{..... (2.9)}
\]

**Hamming Distance**

It is the most popular method to measure the binary attributes. It determines the numbers of bits differ between the two binary strings that many numbers of bits should be changed to convert one string into the other string.

**Levenshtein Distance**

Levenshtein Distance is also known as edit distance, and it is a generalized form of Hamming distance. The distance between two strings is calculated by finding the minimum edit operations needed to change one string into the other. The edit operations can be an insert, delete, or substitution of a single character.

**Soundex Distance**

In this method, the terms are encoded into codes based on their pronunciation. It works according to the phonetic indexing scheme. It increases the efficiency of matches as it accepts the small spelling changes.

**Smith-Waterman-Gotoh Distance**

Smith-Waterman-Gotoh distance method measures the similarity between two strings or vectors. For instance, the two strings are “wwwABCDzzz” and “xxxABCDyyy”. The result of Smith-Waterman-Gotoh distance will be “4”.

**Jaro-Winkler Distance**

The Jaro–Winkler distance is a measure used to measure the similarity between two strings. It returns the similarity score of two strings. It is a complementary method for Levenshtein measure as it gives a score of difference between two strings. It is the most suitable method for short strings like names of persons.
**Overlap Coefficient**

The overlap coefficient is also known as Szymkiewicz-Simpson coefficient. This similarity measurement method measures the overlap between the two sets (A, B). It measures the overlap through the divide, the size of the intersection of the small size of the set of the two sets.

\[
\text{Overlap} (A, B) = \frac{|A \cap B|}{\text{Min}(|A|, |B|)} \quad \ldots \ldots \quad (2.10)
\]

**2.2.2 Structure-Based Similarity Measures**

The structure-based method also known as edge-based method measures the similarity between two terms based on the length that connects the terms and location of the terms in the taxonomy [33].

Path Length Approach: The path length method is a conventional approach to measuring the semantic similarity. It is further divided into the shortest path length and weighted shortest path.

Shortest Path Length: Shortest Path Length presents a straightforward method to measure semantic similarity. This method computes the similarity via calculating the number of edges between the nodes. The shorter length denotes the high similarity. Moreover, this approach provides a good result, but it does not consider the Is-A relationship.

Weighted Shortest Path Length: The weighted shortest path measure is a simplification of the shortest path length approach, and it overcomes the problems found in the shortest path length. This method assigns the weight to every edge and calculates the semantic similarity based on the weight.

**2.2.3 Feature-Based Similarity Measures**

Feature-based similarity measures are alternative to the distance based similarity measure [34]. It measures the similarity based on the shared features of the terms. If
The entities share standard features, it is considered as the similar otherwise they are dissimilar. It defines the following formula to determine the similarity.

\[ S(X, Y) = \alpha g(A \cap B) - \beta g(A - B) - \gamma (B - A) \] ...

Where \( X, Y = \) Entities

\( g(A \cap B) = \) represents the common features

\( g(A - B) = \) represents the distinctive features in A

\( g(B - A) = \) represents the distinctive features in B

\( \alpha, \beta, \gamma = \) represents the weight.

**Tversky Measure**

This measure considers the term features and eliminates the position and information content of the term to estimate the similarity between different concept/terms. Every term describes with its set of features. The general features of the concepts increase the similarity, and unique features decrease the similarity [34]. This measure uses the following equation to estimate the similarity.

\[ \text{Sim}(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_1 \cap C_2| + \propto |C_1 - C_2| + (\propto - 1)|C_2 - C_1|} \] ...

Where, \( C_1 \) and \( C_2 \) represent the feature sets of two terms \( t_1 \) and \( t_2 \) respectively, and \( \propto [0, 1] \). The value of \( \propto \) increased with the common features and decreases with the different features.

**2.2.4 Probabilistic Similarity Measures**

The application consists of complex data, thus, makes similarity measurement as difficult. Finding the accurate feature and position for semantic relation in a complex data, is not an easy task. Example applications are image retrieval, face recognition, DNA analysis and, multimedia databases. Probabilistic similarity measures are introduced to compute the similarity between these types of complex data [35]. This measure exploits the probability density functions to represent the
probability of certain feature values. However, the complexity of the data increases, the computation becomes more difficult.

**Maximum Likelihood Estimation (MLE)**

Maximum likelihood estimation is an approach used to fit a mathematical model to given data. The real world data is modeled using Maximum likelihood estimation that provides a method to change the free parameters of the model. Communication systems, Structural question modeling, Data modeling in nuclear and particle physics, Computational phylogenetics are some of the applications uses the MLE measure.

**Maximum a Posteriori (MAP) Estimation**

MAP estimation is more related to the MLE, but the MLE uses only the experimental measures of data for the similarity calculation. Unlike MLE method, MAP uses the existing distribution for measuring the similarity, and it is a Bayesian approach. However, it is not used in conventional methods due to its complexity and unreliable previous information.

**2.2.5 Hybrid Measures**

This measure derives and combines the knowledge from multiple sources. If one source is inadequate in the knowledge, additional information can be accessed from the alternate data source.

**Li Measure**

This measure combines the shortest path (SP) information and depth of the knowledge source (N) to find the similarity between the concepts [33].

\[
\text{Sim}(C1, C2) = e^{-\alpha \text{SP}} \times \frac{e^{\beta \cdot N} - e^{-\beta \cdot N}}{e^{\beta \cdot N} + e^{-\beta \cdot N}} \quad \ldots \ldots (2.13)
\]

Where, \( \alpha \) and \( \beta \) are parameters scaling the contribution of the shortest path length and depth respectively.
2.2.6 Extended/Additional Measures

Identifying the relationship between the data, represented as a graph is an easiest and efficient way. Various similarity measures are applied to match the graph. Moreover, the similarity measure works on the basis of graph theory and applied to many useful tasks such as computer vision, audio content analysis and retrieval and document structure analysis [36]. In graph match, it does not need the exact match and hence, it presents the error-tolerant matching to handle the noise. Graph, subgraph isomorphism, and maximum common subgraph are considered as the core methods to compute the graph. Some of the similarity measures apply the similarity on the basis of the fuzzy set theory. These measures are working based on union and intersection operations, maximum difference, and on the differences and summation of set membership values.

2.3 Sources of Semantic Similarity

The sources of semantic similarity are the concept features, word associations, word co-occurrence or lexical context and semantic relation [37].

2.3.1 Features

The feature is the first and most attractive source of semantic similarity. It denotes the physical, functional and characteristics of the particular entity. Features have a significant impact on semantic similarity determination. To obtain a similarity between two concepts, examine the feature set of both concepts. The more common features between the concepts tend to increase the similarity, and non-common features increase the dissimilarity. The semantic similarity measures use the ontology, corpus and dictionary to identify the common characteristics of the objects. These sources produced a better result in the following tasks: response times in a speeded categorization task, typicality [38], similarity judgments [39] and concept coherence [40].
2.3.2 Association

The important concept to discover and rank the similarity between the objects is an association. It is the source that determines the statistical strength of the relationship between the concepts. It is defined as “An association presence between the two variables if the probability of occurrence of one variable depends on another variable”. The association between the concepts can be Positive/ Negative, Direct/ indirect, or Casual/Formal. The association can be measured using the Dice coefficient, Overlap ratio, Jaccard measurement, and Pointwise Mutual Measurement.

2.3.3 Lexical Context

Lexical context denotes the linguistic factors that influence the meaning of the text. It is the probability of the co-occurrence or collocation of one lexical unit with the other, which forms part of the meaning. The meaning of the word indicates the context in which a particular word is being used. It assumes that the phrases contain similar meaning or relationship occurs within a context. Lexical context is useful to word sense disambiguation.

2.3.4 Semantic Relation

The semantic relation is the final source of similarity that finds the similarity between the concepts to estimate the similarity. The relationship between the concepts can be derived from any of the structured information sources like ontology. One of the well known knowledge sources for semantic similarity is WordNet. WordNet is a lexical database that groups the synonyms with semantic relationships, such as Hypernyms, Hyponyms, Holonyms, and Meronyms. The measure of the semantic relation between the concepts has a significant impact on text processing applications.
2.4 Need for Ontology Extension

Ontology is the popular source among all other sources to find the semantic similarity between the two concepts/terms [10]. Ontology is a well-defined formal logical model to represent the knowledge that describes the facts through defining classes, relationship, and attributes. It works as a supporting tool to find the similarity between the terms. In ontology, words are defined as linguistic objects in a lexical or terminological database and these are connected through the different relationship in a hierarchical manner. However, the ontology lacks in deriving at new words, and the accuracy of the relationship is a primary problem of ontology [26].

2.4.1 WordNet

WordNet is a manually compiled online electronic dictionary organized as a graph [19]. It is domain independent since it does not cover any domain specific information. WordNet comprises of general nouns, verbs, adjectives and adverbs, organized in terms of their meanings around lexical-semantic relations, which include synonymy and antonymy, hypernymy and hyponymy, meronymy and holonymy. It groups the words that have similar meanings called synset. The synset connects with the other synset using any one of the above relationships. The latest version of WordNet is 3.1 which contains 155,287 words organized in 117,659 synsets for a total of 206,941 word-sense pairs. A typical application of WordNet is to find the semantic similarity between the words. The similarity measures estimate the distance between the words in a WordNet hierarchical organization through counting the number of edges between the synset which contains the words. The computed distance indicates the degree of the similarity among the given words.

WordNet does not contain the domain-specific information. Hence, it is not applicable for domain specific applications. Moreover, many useful new relationships among the concepts arise over the time due to the prolific use of English are found missing in WordNet. Thus, the proliferation of senses and lack of new evolving relations between concepts in WordNet tend to reduce the
performance of the applications. Hence, WordNet should be updated with the newly arrived relationship derived from any of the updated knowledge sources.

2.4.2 Wikipedia

Wikipedia is a freely available multilingual online encyclopedia, and anonymous volunteers update the online information every day. Any volunteer can create or edit the Wikipedia page and its contents. Due to the collaborative efforts of the volunteers, the number of articles in the Wikipedia is growing rapidly. From February 2001 to till date the number of English articles has increased from 1,000 to 3 million. Sometimes it may affect the quality of the articles. However, the fruitful advantage to overcome this drawback is the collaborative nature of Wikipedia. Collaborative nature indicates that if any error occurs in the article, any volunteer can correct the content. Moreover, if any user involved in deliberate vandalism, the particular user can be detected and blocked by an administrator. Every article in Wikipedia comprises a single web page, and it describes a single topic. Articles are hypertext documents, and it connects to the other related articles that can present in both inside or outside of Wikipedia. One of the essential content of Wikipedia is the disambiguation page. It comprises of a list of links to articles that contain different meanings for the same name. It denotes the word disambiguation in the title or the tag. Moreover, it is the inevitable source in NLP applications like relation detection and disambiguation [24] [27] [28].

2.4.3 Web Search Engine

Web search engine provides an efficient way to access the needed information among the vast amount of information on the web. The Web is a colossal knowledge base comprises of updated information and hence, it is the most suitable to measure the similarity between the words/concepts. Web search engines facilitate Page counts, and snippets that are useful information sources to support similarity [23] [25]. Page count is the result of some pages comprising the query words. For instance, the words “Apple” and “Computer” are given as input to the web search engine and obtain an output 34,20,00,000 results. While the query of “Banana” and
“Computer” returns 2,24,00,000 result. The result gives evidence that the Apple and computer are more similar than the banana and computer. The snippet is the short descriptions of text returned by the search engine appear in the every search result with the query term. It also additionally helps to identify the semantic similarity between the texts.

2.5 Information Retrieval Models

Retrieving relevant information for the user needs from this vast information is known as “Information Retrieval”. Similarity measurement helps to increase the quality of information retrieval and reduces the information processing time and cost. Several similarity measures are proposed and used by different applications. The IR system exploits the defined similarity measures to select a relevant result and to present a ranked relevant result. Following are some of the IR models [41]:

2.5.1 Set-Theoretic Models

This model processes the information according to the set theory, and similarities are derived using the set-theoretic operations.

**Boolean Model:** The straightforward IR model is a Boolean model, and it works on the basis of the set theory. It performs an exact match to retrieve the result of the user query and performs operations such as AND, OR and NOT. The advantages of the Boolean IR model are: simple implementation and efficient computation. However, disagree on an exact match, complex query construction, unavailability of ranked results are considered as the drawbacks of Boolean model.

**Fuzzy Set-Based Model:** The fuzzy set based approach is proposed as an IR model to overcome the limitations of the Boolean model. In this model, the degree of membership of elements is varying. Mixed Min and Max and the Paice model are the types of Fuzzy set-based model. Both models use the P-norms algorithm to compute the query weights. In fuzzy-set theory, an element has a varying degree of membership [42].
Mixed Min and Max Model: In this approach every index term associated with its fuzzy set. The degree of relationship of the document is known as document’s weight, and it is represented as an index term. The degree of the relationship is defined as:

\[ D_{X \cap Y} = \min(d_X, d_Y) \] .......... (2.14)
\[ D_{X \cup Y} = \max(d_X, d_Y) \] .......... (2.15)

Paice Model

The Paice model extends the MMM model to support every term weights to compute the similarity between the MMM model uses only the minimum and maximum weights for the index terms.

\[ S(D, Q) = \sum_{i=1}^{n} \frac{r^{i-1} \ast W_{di}}{\sum_{j=1}^{n} r^{j-1}} \] .......... (2.16)

Where r is a constant coefficient and \( W_{di} \) indicates that it is arranged in ascending order for ‘and’ queries and descending order for ‘or’ queries.

Extended Boolean Model

The extended Boolean model performs better than other models as it provides weight and terms position to the terms [43]. The term weight assists to generate the result in the ranked order. It overcomes the limitation of Boolean model and structural characterization problem in the vector model.

2.5.2 Algebraic Models

IR models are working on the basis of algebraic calculation called as Algebraic models. In this approach, the terms are represented as vectors/matrices.
2.5.2.1 Vector Space Model

The vector space is an algebraic model used in information filtering, information retrieval, indexing, and relevancy ranking. In this model, the text documents and the query terms are represented as vectors. If the term occurs in the document, the value of the term is non-zero in a vector. The weight of the vector can be computed in different ways. The concept term frequency and inverse document frequency (tf-idf) are the familiar and a favorite way to calculate the weight of the term [44]. The term can be a single word, keyword or a phrase. The similarity between the document vector and query vector is computed using the cosine similarity measure. The query terms match with the document collections to retrieve a relevant document using these vector operations [45]. It can be calculated by matching the deviation of angles between the query vector and the document vector. If the result of the cosine similarity is zero, it indicates the document and query vector are orthogonal it means the query term is not present in the document.

\[
\cos \theta = \frac{d_2 \cdot q}{\|d_2\| \|q\|} \quad \ldots \ldots (2.17)
\]

Where \(d_2\) = document vector, \(q\) = query vector,

\[
\|d_2\| = \text{norm of vector } d_2, \|q\| = \text{norm of vector } q
\]

The norm of the vector is estimated using the following equation

\[
\|q\| = \sqrt{\sum_{i=1}^{n} q_i^2} \quad \ldots \ldots (2.18)
\]

2.5.2.2 Semantic Similarity Retrieval Model (SSRM)

The Semantic Similarity Retrieval Model (SSRM) measures the document similarity that assists the information retrieval model to retrieve an appropriate document among the large amount of documents for the user query [12]. In this model, user query and documents are represented as a term vector. The weight
indicates the number of frequency of the term in the collection of documents that is calculated through different ways. The term frequency - inverse document frequency (tf-idf) model is one of the ways of computing the weight. It calculates the document and vector similarity using the cosine similarity. To overcome the limitation, SSRM includes the feature to discover the semantically similar terms for the query and document terms using WordNet and semantic similarity methods. SSRM comprises the following techniques to include the semantic similarity.

**Term Re-Weighting**

Term re-weighting helps to adjust the weight of the terms based on its relatedness with the similar terms in the same vector. Hence, the terms that are semantically similar obtain higher weight than the non-similar terms. The following formula finds the re-weighting of the terms.

\[
q_i = q_i + \sum_{\text{sim}(i,j) \geq t} q_j \cdot \text{sim}(i,j)
\]

... ... (2.19)

Where \( t = \) user defined threshold, \( i,j = \) query terms, \( q_i, q_j = \) weight of the query terms.

**Term Expansion**

The term expansion assists to expand the query terms with the help of semantically similar terms. Initially, it selects the most common synonym terms for the query terms. Consequently, the other terms semantically related to the hyponyms and Hyponyms of the query terms are also extracted. Every query terms are represented by using the WordNet hierarchy. The terms near the query terms are examined using the following equation and the terms that have similarity more than 0.9 are included in the query vector.

\[
q_i = \begin{cases} 
\sum_{\text{sim}(i,j) \geq T_n} \frac{1}{n} q_j \cdot \text{sim}(i,j), & \text{if } i \text{ is a new term} \\
q_i + \sum_{\text{sim}(i,j) \geq T_n} \frac{1}{n} q_j \cdot \text{sim}(i,j), & \text{if } i \text{ had a weight } q_i
\end{cases}
\]

... (2.20)

Where, ‘n’ is the number of hyponyms of each expanded term j.
Document Similarity

Document similarity finds the similarity between an expanded and re-weighted query and a document using the following equation. According to this result, the document will be retrieved for the user query.

$$\text{Sim}(q, d) = \frac{\sum_i \sum_j q_i d_j \text{sim}(i, j)}{\sum_i \sum_j q_i d_j}$$  \hspace{1cm} \hspace{1cm} \text{(2.21)}$$

Where $q_i$ = query term, $d_j$ = document term

2.5.3 Latent Semantic Analysis - Based Model and Neural Networks

The vector space model constructs the vector based on the string terms and its frequencies, and it performs the accurate string matching in information retrieval. Hence, it retrieves only the syntactically similar concepts and not considers the semantic similarity. The Latent Semantic Analysis technique was proposed to cross over this limitation and increase the information retrieval accuracy. Latent Semantic Analysis is the method uses the Singular Value Decomposition (SVD) to convert the document to a term vector [46]. Moreover, it effectively handles the relationship, such as synonymy and polysemy between the concepts. A neural network is a traditional machine learning approach that efficiently extracts the information from the data [47]. Spreading activation supports the neural networks for information learning and intelligent matching using the data represented as a weighted and interconnected graph.

2.5.4 Probabilistic Models

These approaches retrieve the information according to the probability of a relationship between the data. Hence, these techniques are applied with information uncertainty. Inference network explicitly represents the probabilistic inferences, and it combines multiple pieces of evidence to compute the conditional probability. Moreover, it generalizes the other models [48]. Probabilistic models comprise the document network and query network, and it captures the significant dependencies among these networks. A Bayesian network is a widely probabilistic graph approach.
2.5.5 Knowledge-Based and Structure-Based Models

In the knowledge-based models, the information retrieval uses any of the knowledge sources to determine the semantic relationship between the concepts/terms. Most of the semantic similarity approaches use the knowledge sources as the base to estimate the similarity [49] [50]. In general, the conventional IR methods are more valuable to the data contents and do not consider the structural information about data. With the intention of improving the IR performance, the structure-based models combine the content and structural characteristics of the data [51].

2.6 Applications Based on Semantic Similarity Approaches

Similarity measures play an increasingly important role in text related research and applications. This section discusses the existing application based semantic similarity approaches.

2.6.1 Information Retrieval

The system implements the novel information retrieval method called Semantic Similarity-based Retrieval Model (SSRM) to measure the similarity between the documents. This method is capable of determining the similarity between the documents containing conceptually similar terms. The semantic similarity approach implements a concept of similarity matching method based on information content using the hierarchy of WordNet [52]. This method finds the similarity between the words. It provides a search engine using the Google API and expands the each term of user’s queries with a set of synonyms using the similarity score of user’s queries to enhance the IR.
2.6.2 Data Mining

Data mining performs the following: i) Derives useful information from the implicit and hidden information. ii) Describes the meaningful patterns from the massive amount of raw data collections. The similarity measure between two information concepts/terms is a fundamental need for the both information retrieval and Data mining. The following are some of the data mining applications that employ the semantic similarity measures.

2.6.2.1 Classification

Both the clustering and classification techniques help to identify the hidden patterns in data. However, both use the semantic similarity measures to find the similarity between the objects. A novel classification algorithm is proposed to classify the documents according to the word meanings and the relationships [53]. The proposed approach is developed based on the word structures provided by WordNet. WordNet arranges the words into groups of synonyms called Synsets and arranges the Synsets into hierarchies to represent the relationships between concepts. A semantic similarity method incorporates the WordNet knowledge into a text representation [54]. This approach significantly reduces the error rate in particular types of text classification tasks. It exploits the lexical and semantic knowledge in WordNet. It represents the text based on the WordNet hierarchies. Rules are framed by the commonly used bag-of-words representation.

2.6.2.2 Clustering

Clustering is the process of grouping the objects that have the similar features in one group and the dissimilar objects in another group. Initially, the data set is divided into groups according to the semantic similarity between the data. Consequently label will be assigned to the each group. It quickly adapts to the changes. K-means clustering, and expectation maximization (EM) clustering are the most widely used clustering techniques [55]. Coupled clustering is a framework that assists to reveal the analogies between sub-structures of distinct composite systems [56]. The method applies to the synthetic as well as textual data. The semantic
similarity measure can also estimate the semantic similarity between the words using its distribution in a corpus [57]. It gives the ranked list of similar words to each word as output. Moreover, using this measure it constructs the thesaurus using a parsed corpus and it presents the methodology to evaluate the automatically constructed thesaurus. Finally, the system validates the proposed system similarity analysis with the Euclidean (centroids) distances and Pearson correlation coefficient using the Iris dataset [58].

2.6.3 Natural Language Processing

Natural Language Processing (NLP) is a component of artificial intelligence (AI). It is a technique translates the human language into the machine understandable language [59]. Word sense disambiguation, and named entity recognition are some important applications of NLP that uses the semantic similarity. Word sense disambiguation is the process of identifying the sense of a word in a particular context when it has several meanings while extracting the possible senses of a word. In traditional approaches, the potential senses are derived from the dictionary, and now the modern methods use the WordNet. All the derived possible senses match with the particular context. The degree of similarity between the potential senses and the specific context helps to decide the sense of a polysemic word.

2.6.3.1 Named Entity Disambiguation

Named entity disambiguation aims at addressing the problem of providing sense to the proper name in a text. The novel named entity disambiguation approach associates the words representing Named Entities with their identities using the semantic relatedness scores secured with a graph-based model over Wikipedia. This approach considers only the named entities within the text and does not make the bag of Words representation of text for disambiguation. Another approach performs the named entity disambiguation using the semantic similarity measure. It exploits the Wikipedia semantics to build a disambiguation dictionary and vector–based word model [60]. Through the explicit semantic analysis, the analyzed documents
are converted into semantic vectors. The experimental result shows that the proposed approach performs better than the traditional approaches.

2.6.3.2 Question Answering

The task of computer assisted assessment of short student answers is discussed in [61]. The approach combines the graph alignment features with lexical semantic similarity measures using machine learning techniques. The DIOGENE is the question/answering system that exploits three components such as question processing, search and answer processing component [62]. The question processing component works based on the several linguistic processors and resources including WordNet. The search component is based on information retrieval techniques and answer processing component exploits the similarity between the question and the documents to identify the correct answer. Unlike the automatic scoring systems, the proposed method uses the unsupervised approach for the short answer grading task using text to text similarity [63].

2.6.3.3 Word Sense Disambiguation

The disambiguation method called “Weighted Overlapping” is implemented to extend the Lesk’s approach for disambiguation [64]. For an accurate word “Synset”, “Hypernymy” relation, and definition of the context features are extracted from the WordNet database and used as an input to the Disambiguation algorithm. Using this WordNet definition for each sense of the word a sense bag is created that forms the context bag. According to the related synset’s position in the WordNet taxonomy, assigns the weight to every word in the context bag. The disambiguation of a word in a context is calculated by comparing the similarity between the words of the sense bags and the context bag. The unsupervised graph-based algorithm helps in word sense disambiguation [65] using the identified similarities between word senses and centrality algorithms. The proposed algorithm annotates all the words in a text.
2.7 Summary

This chapter briefs the basics of semantic similarity and the classification of semantic similarity measures. Outlines the details of the various similarity measures and discusses the sources involved in semantic similarity. Finally, it examines the semantic similarity approaches used in different applications.