Investigating the Similarity of Court Decisions

Sarika Jain[1], Deepak Jaglan[2,*] and Kapil Gupta[3]

National Institute of Technology Kurukshetra, Department of Computer Applications, Kurukshetra, India

Abstract

The association between words, phrases, and documents is referred to as semantic similarity. Semantic similarity has played a significant role in internet search engines regarding content ranking. It also has wide applications in information retrieval, artificial intelligence, etc., to name a few. This paper reviews the general architecture, categorization of approaches, and techniques and metrics for determining semantic similarity between documents in a comprehensive way. We have conducted experiments on the different statistical methods, viz., word vector-based techniques (TF-IDF, LDA, Word2Vec, Doc2Vec, Glove, and fastText), and transformer-based techniques (Longformer-base, Sentence-BERT-large-nli, Sentence-BERT-large-nli-stsb, and Sentence-RoBERTa-large-nli-stsb) over Indian Supreme Court decisions and discussed the results. The Doc2Vec approach over the whole document is found to correlate the most with the expert judgment.

Keywords

Semantic Similarity, Legal Document, Document Embedding, Cosine Similarity

1. Introduction

Semantic similarity can be well described as the relate-ability between words, sentences, and documents. It is most likely a quantitative measure of information that has evolved into a core technique that is now widely used in a variety of fields, including biological computing [1], information retrieval [2], artificial intelligence [3], geoinformation [4], and natural language processing [5], as well as other intelligent knowledge-based systems [6]. For the use case scenario, identification of related literature assists legal professionals in obtaining relevant literature. Some authors have studied similarity analysis of legal judgements [7]. We bring relevant literature primarily based on text-based methods and deep learning approaches like transformer models.

Our focus in this paper is to review the semantic similarity approaches exhaustively in context to the legal case documents in particular. This approach is not restrictive to the legal case documents. Instead, we may use this method in various other subjects' domains. Throughout this article, we shall concentrate on the legal arena.

The requirement for an accurate and relatable legal information retrieval area is the most pressing challenge in today's legal society. Because the Common Law System is one of the most widely followed legal systems globally, the success or failure of a case is heavily influenced by

ACI'22: Workshop on Advances in Computation Intelligence, its Concepts Applications at ISIC 2022, May 17-19, Savannah, United States

*Corresponding author.

jasarika@nitkkr.ac.in (S. Jain); deepakjaglan34@gmail.com (D. Jaglan); kapil@nitkkr.ac.in (K. Gupta)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

previous instances. The deluge of information on the internet has made it difficult for legal practitioners to manually discover significant earlier examples that appropriately serve their current case. As a result, a likely answer is found by comparing the similarity of the different case documents, which various authors have recently studied [8, 9, 10]. Statistical methods, also known as text-based methods, utilize the textual content of legal documents. These methods include only primitive text-based similarity measures, such as TF-IDF-based approaches. In [10], the authors have improved the text-based technique with the similarity measures such as topic modeling and neural network models such as word embeddings and document embeddings. Also, it has been shown that the word vector-based approaches perform better than other approaches.

The present paper details the outlines of the review of different methods used in the text-based category besides approximate validation of the experimental results. More precisely, we discuss:

1. The comprehensive details of the different semantic similarity approaches provide an insight into the generalized architecture of the various techniques used in semantic similarity.
2. In context to the legal domain, we confine ourselves to word vector-based and transformer-based approaches and discuss the experimental results we obtain in each method in both directions.

The layout of this paper is as follows: In the next section, we present the analytical discussion comprising the general architecture for semantic similarity along with the different semantic similarity approaches in detail. A brief discussion regarding the similarity measures followed by evaluation measures is required for a comparative study in Section 3. Finally, in the same section, we detail the experimental results obtained in the context of the legal domain, followed by the underlying discussion. Section 4 deals with the conclusions.

2. Analytical Discussion

The main content of the present paper is depicted in the following flowchart (Figure 1), i.e., first of all, we will discuss various document representations methods and document pre-processing. After that, we will discuss the semantic similarity approaches followed by semantic measures and evaluation measures. The meaning of these terminologies shall be transparent in their respective discussions.

General Architecture

We feed the input as the unstructured text, and from it, we select the corpus to obtain the representative text. After that, we initiate the data processing of the representative text by first removing punctuations and stopwords and then stemming. Thus we get a clean text. Now, we begin the data modeling of the clean text to extract the features of documents, i.e., the word embeddings. For that, we employ semantic similarity techniques, viz., word vector-based techniques (TF-IDF, LDA, Word2Vec, Doc2Vec, Glove, and fastText), and transformer-based techniques (Longformer-base, Sentence-BERT-large-nli, Sentence-BERT-large-nli-stsb, and

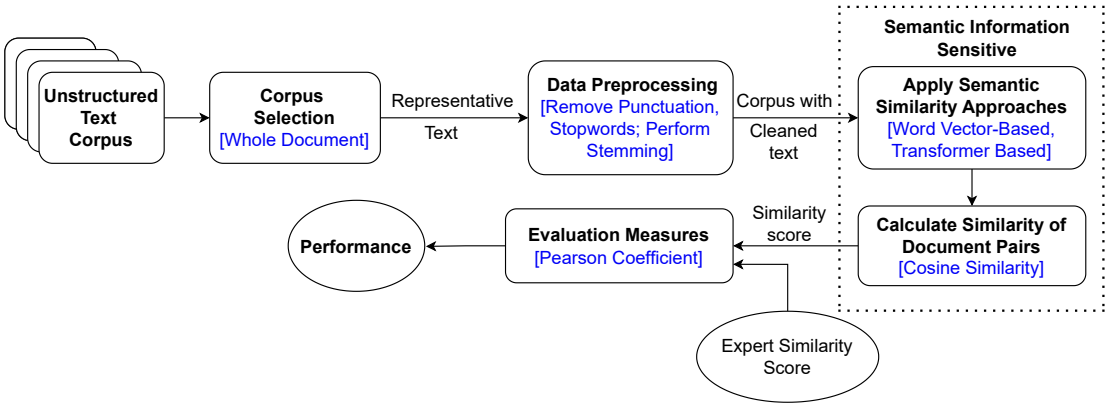

Figure 1: General Architecture for Semantic Similarity.

Sentence-RoBERTa-large-nli-stsb). We calculate the cosine similarity between these feature documents to obtain the similarity scores. Data modeling and similarity measures can utilize semantic information. Further, since we also have the similarity scores from the experts, we will evaluate the Pearson coefficient between the similarity scores given by the expert with the ones obtained by us. Hence, we quantify how well our methods perform when compared to the similarity scores given by the experts.

2.1. Document Representation

There are various ways to document representation, viz., whole document, summary, paragraph, and the reason for citations (RFCs).

In whole document representation, the whole of the document is taken under consideration, while in summary, the important content is taken into consideration, leaving the redundant part. A set of paragraphs is considered in paragraph document representation in such a way that each paragraph of one document is compared to all the paragraphs of the other document in the corpus. RFC method is a citation-based method, and it works on a similar note as the paragraph-based method. In thematic representation, the theme of the document is taken into consideration. After selecting meaningful representations from the text of the documents, their similarity is measured.

2.2. Data Pre-processing

Data preprocessing is crucial in preparing the data since we deal with unstructured text data. It transforms the text into a more digestible form. Now, we outline the steps involved in the data preprocessing. Firstly, all of the letters are changed to small lowercase. Then based on whitespaces, tokenization of the text into words is done. Except for terms containing the letters hyphen, dot, and comma, all non-alphabetic words are filtered away. After that, standard English stopwords are then removed from the list of words. Using Porter Stemmer, we finally perform the overall word stemming. In this way, we obtain a better representation of our text.

2.3. Semantic Similarity Approaches and Measures

The main principles behind the existing approaches that we reproduce in this study are described in this section. As previously indicated, existing approaches utilize various similarity measures that are divided into three broad categories: (i) statistical similarity, (ii) graph-based similarity, and (iii) document clustering-based similarity. We will present a detailed overview of each category classified above.

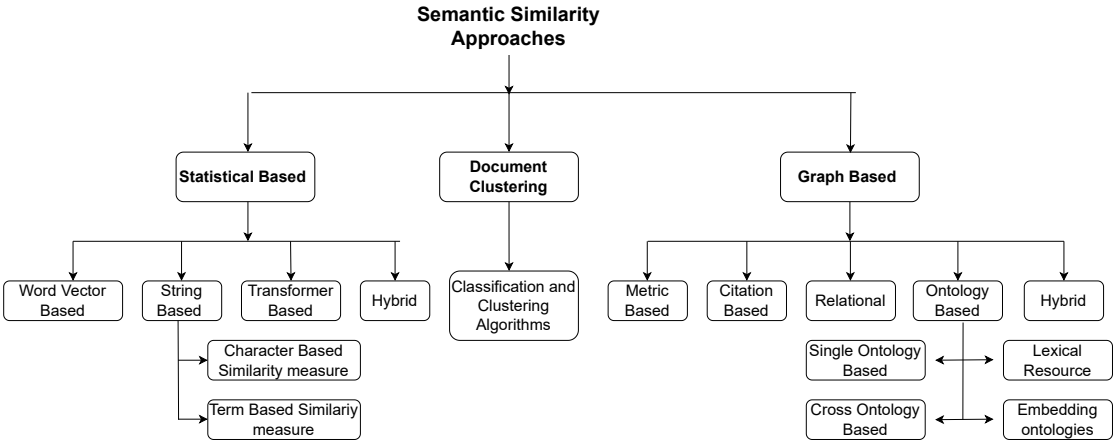

Figure 2: Semantic Similarity Approaches.

2.3.1. Statistical Similarity

The statistical-based similarity approach is built on collecting texts either in written or spoken forms. There are various ways to compare statistical similarities between legal documents, viz., word vector-based, string-based, transformer-based, and hybrid-based. We confine our experiments to the word vector-based and transformer-based techniques in this paper.

The meaning of the word vector-based method is clear from its very name, i.e., it defines the vector representation of the documents. We enlist all the methods derived from word vector, viz., TF-IDF technique, LDA, Word2Vec, Doc2Vec, Glove, fastText. A single vector representation of the given document (e.g., a legal document) is created in the TF-IDF approach. The computation of the similarity score between vectors is done with the aid of the cosine similarity (see, e.g., [8]). In contrast, as depicted in [10], the LDA technique is a topic modeling algorithm, and it captures the semantics of the documents in an appropriate way. In the models, based on neural networks such as Word2vec and Doc2vec, gives a vector for each distinct word (see, e.g., [11]) and each document (see, e.g., [12]), respectively. Similar to word2vec method, the dense vectors are constructed in both these GloVe (see, e.g., [13]) and fastText methods (see, e.g., [14]).

String-based similarity includes the character and term-based similarity measures. The transformer-based similarity approach is built on the language models with deep contextual text representations by incorporating the word positioning. The various transformer techniques

are given as Longformer-base, Sentence-BERT-large-nli, Sentence-BERT-large-nli-stsb, and Sentence-RoBERTa-large-nli-stsb.

To address the constraints of the numerous statistically-based similarity approaches listed above, a hybrid model was created by combining some or all of them in a suitable way to meet at least all of the essential criteria of each feasible combination of methods. For more details in the context of the hybrid method, the reader is referred to [15].

2.3.2. Document Clustering

Clustering is an unsupervised learning problem in which the goal is to arrange a set of objects in such a way that the objects in the same cluster are more similar (in meaning) to each other than the objects in the other cluster. Clustering may be used in various disciplines, with intelligent text clustering being one of the most common. Traditional text clustering algorithms gathered documents based on keyword matching, which meant that the texts were grouped without any descriptive concepts. As a result, non-similar texts were grouped. The essential answer to this challenge is group papers based on semantic similarity, which means grouping pages based on meaning rather than keywords.

One of the most well-known methods for producing a single grouping is k-means, wherein the number of clusters, k, must be determined beforehand. Initially, there are k clusters specified, and after that, each document in the document collection is reassigned based on the document's resemblance to the k clusters. The k clusters are then updated. After that, the document set's documents are all reassigned. This method is repeated until all k clusters remain the same. Alternatively, from [16], bisecting k-means method is used to cluster documents. Here, all items are thought to be part of a single cluster. A cluster is broken into two every time. This process is continued until the desired number of clusters has been achieved. The reader is referred to [17, 18, 19] for more details on clustering approaches.

2.3.3. Graph Based Similarity

The graph-based similarity approach is based on graphical methods. These methods are further based on different techniques, ontology-based, relational-based, citation-based, and hybrid-based. The prior-case citation network of the document is constructed to compute the Precedent Citation Similarity. The vertices of the network are the case documents. A directed edge exists between two vertices i and j if document i cites document j in its text. Consider an example graph such that an edge exists from vertex A to E since A cites E. To build document vectors, we investigate citation-based networks approaches in which documents are nodes and edges correspond to citations.

The relational approach emphasizes measuring the relation between two words, unlike measuring the degree of similarity. Using a predetermined pattern of vector frequencies from a vast corpus, this approach determines the link between word pairs. It enhances current ontologies and is utilized in document semantic annotation. The reader is referred to [20, 21, 22] for more details on these three approaches.

The ontology-based approach is a graph-based semantic similarity approach, and it is

classified into three broad methods: single ontology-based, cross ontology-based, and lexical resource. The path distance between concepts determines how similar the two concepts are. The ontology or taxonomy structure is used to calculate similarity in this metric. A type relation links essential linkages in this ontology or taxonomic structure. As a result, the shortest path is used to compute similarity, and the length of the path defines the degree of similarity. The depth relative measure is similar to the shortest path approach, but it takes into account the depth of the edges linking the two concepts in the ontology's basic structure and determines the depth between the root and the target concept. In the information-based approach, also known as the corpus-based approach, the information previously contained in the ontologies or taxonomy is supplemented with the knowledge given by the corpus. For comparing the concepts, the hybrid and feature-based measures consider the knowledge derived from different sources and features, respectively. We refer the reader to [23] for further details on the DeepWalk algorithm.

Previously mentioned semantic similarity measurements are intended for a single ontology. With the expansion of online information sources, metrics are needed to calculate the similarity between concepts belonging to different ontologies. The methods that quantify the comparison of the terms from various ontologies are known as cross ontology measures.

To compute the semantic similarity, one employs WordNet and Wikipedia as Lexical resources. The wordNet technique is based on Directed Acyclic Graphs (DAG) theory. The semantic distance and DAG information compute the semantic similarity between the words or concepts. We refer the reader to [24] and [25] for further details on DAG.

The hybrid methods can be a combination of statistical, ontology, and relational approaches. We refer the reader to [26] for more details on such approaches.

3. Experimental Results and Discussion

This section compares these scores to those assigned by domain experts to see if they are consistent. We have taken the data sets of legal documents, viz., Indian Supreme Court case decisions (gold standard pairs) (see 3.1), for legal document similarity.

3.1. Dataset

The dataset contains all Indian Supreme Court case decisions in text format spanning 67 years (from 1950 to 2016). Each text begins with an optional headnote (a summary of a legal case that incorporates several legal concerns and specifies the written laws employed throughout the litigation process) and continues with the case's whole litigation procedure. We crawled the texts from the Legal Information Institute of India's (LIIofIndia) website (http://www.liiofindia.org/ in/cases/cen/INSC/), a website that maintains several legal databases. A gold standard comprising legal expert judgments on how similar two documents are, is essential to compare and evaluate our methods. We have analyzed the 47 pairs of the case documents of the Indian Supreme Court, as our gold standard, along the lines of [8] and [10]. The expert annotations ranging from 0 (lowest similarity) to 10 (highest similarity) were sought for each of these pairs.

3.2. Evaluation Measure

We calculate the similarity scores using each of our techniques for each of the 47 test pairings to assess our techniques. Then, for each strategy, we find the Pearson Correlation Coefficient between the 47 scores obtained by the techniques to those provided by the experts.

3.2.1. Calculate Similarity between pairs

Finding similarities among documents is vital from the perspective of Information Retrieval and allied fields. The approaches create for two documents a vector representation with the dimensions being the terms in the documents, word embeddings, or semantic notions. As a result, we obtained the vectors of the document pairs. Finally, we apply cosine similarity to find the angle between the resultant vectors.

Cosine Similarity: It is a similarity measure of two non-zero vectors of an inner product space, which finds the cosine of the angle between them. The cosine similarity of two vectors having the same orientation is 1, and vectors that are orthogonal have the similarity of 0.The cosine similarity $\cos(\theta)$ of two vectors A and B is

$$\cos(\theta) = \frac{A \cdot B}{||A||\ ||B||},$$

where, $(\cdot)$ represents the vector dot product.

3.2.2. Performance

The Pearson's correlation coefficient is used to measure how well our approaches work compared to expert similarity scores. The correlation between the obtained scores and those offered by legal experts is then calculated.

Correlation coefficient (ρ)*:* It is the ratio of the two variables' covariance and standard deviations. Mathematically, let P and Q be two variables, then correlation coefficient (ρ) is defined below as

$$\rho = \frac{cov(P, Q)}{\sigma_P \sigma_Q},$$

where $cov(P, Q)$ represents the covariance between P and Q, and σ_P and σ_Q represent the standard deviations of variables P and Q. Also, we have the inequality that $-1 \leq \rho \leq 1$. The value $\rho = -1$ signifies that the variables are anti-correlated whereas $\rho = 1$ signifies that they are highly correlated.

3.3. Results and Discussion

Table (1) enlists the column headings, viz., Case Pairs, Expert Scores, and Similarity scores using word vector-based and transformer-based approaches. This table shows different similarity

scores-given (1) by legal experts and (2) by those that we obtained from the experiment by using word vector-based and transformer-based techniques. To find the similarity scores between the pairs, we used the case of word vector-based the following approaches: TFIDF, Doc2vec, GloVe, and fastText methods, while in transformer-based, we employed Longformer-base, Sentence-BERT-large-nli-stsb, and Sentence-RoBERTa-large-nli-stsb.

In Table (2), for each technique, viz., word vector based and transformer-based, we compute the Pearson correlation coefficient for each method with respect to the expert scores. The highest correlation value obtained for both the approaches is in the italic font, i.e., Doc2Vec and Sentence-RoBERTa-large-nli-stsb.

The methods corresponding to which the detailed similarity scores between each pair are computed in the Table (1) are represented by the bold font in the Table (2). When the expert scores are assigned low, the word vector-based technique is closer to the expert scores than the transformer-based technique. Whereas, when the expert scores are assigned as high, the transformer-based approach is closer to the expert scores than the word vector-based. The Pearson correlation coefficient in the transformer-based method is lesser than that of the word vector-based. This trend can also be seen in [15] where the authors obtain that the evaluation parameters are lesser in transformer-based methods as compared to word vector-based methods in the context of the US Supreme Court decisions. The higher the value of the correlation, the better the corresponding method's performance. Doc2vec obtains the highest correlation value with the experts' score (is computed as 0.685) and Sentence-RoBERTa-large-nli-stsb (is computed as 0.401) methods, respectively, in the word vector-based technique and the transformer-based technique. Overall, Doc2vec provides the highest correlation value with the experts' scores.

4. Conclusion

This paper presents a comprehensive review of the semantic similarity, i.e., categorization, and techniques and metrics for determining semantic similarity. We then discuss exclusively the semantic similarity of the legal court case documents wherein we confine ourselves to word-vector-based and transformer-based techniques in the context of the experiments. Finally, we discuss the results we obtained while computing semantic similarity among legal documents with different techniques, viz., word vector-based techniques (TF-IDF, LDA, Word2Vec, Doc2Vec, Glove, and fastText), and transformer-based techniques (Longformer-base, Sentence-BERT-large-nli, Sentence-BERT-large-nli-stsb, and Sentence-RoBERTa-large-nli-stsb). We observed that the Doc2vec similarity correlates the most with expert judgment from both the techniques, viz., word vector-based and transformer-based techniques.

5. Acknowledgment

This work is supported by the IHUB-ANUBHUTI-IIITD FOUNDATION set up under the NM-ICPS scheme of the Department of Science and Technology, India.

Table 1
Similarity scores for all the gold standard pairs of Indian supreme Court decisions, using word vector base and transformer based approaches.

Sr. No.	Case Pairs	Expert Score	Similarity Scores						
			Word Vector Based				Transformer Based		
			TFIDF	Doc2vec	GloVe	fastText	Longformer	BERT	RoBERTa
1	1992_47 & 1992_76	0	0.168	0.160	0.864	0.347	0.993	0.443	0.351
2	1992_76 & 1992_182	0	0.143	0.146	0.838	0.386	0.993	0.449	0.378
3	1972_11 & 1984_115	0	0.127	0.084	0.838	0.179	0.986	0.643	0.466
4	1969_57 & 1980_91	0	0.282	0.271	0.910	0.521	0.989	0.565	0.625
5	1959_151 & 1982_28	0	0.237	0.238	0.895	0.527	0.990	0.677	0.492
6	1976_200 & 1959_151	0	0.218	0.051	0.904	0.304	0.990	0.674	0.424
7	1985_114 & 1959_151	0	0.291	0.263	0.899	0.572	0.990	0.777	0.683
8	1966_236 & 1967_267	0	0.236	0.353	0.903	0.688	0.983	0.563	0.611
9	1961_34 & 1979_110	0	0.303	0.322	0.935	0.635	0.995	0.689	0.556
10	1961_34 & 1987_37	0	0.151	0.193	0.885	0.447	0.992	0.671	0.615
11	1992_47 & 1987_315	0	0.388	0.358	0.898	0.712	0.991	0.689	0.613
12	1984_115 & 1987_315	0	0.489	0.459	0.960	0.796	0.991	0.712	0.498
13	1992_47 & 1992_76	0	0.168	0.160	0.864	0.347	0.993	0.443	0.351
14	1984_115 & 1987_315	0	0.246	0.238	0.842	0.502	0.990	0.708	0.556
15	1983_129 & 1983_27	1	0.590	0.561	0.959	0.723	0.995	0.745	0.657
16	1979_110 & 1953_28	2	0.481	0.178	0.957	0.583	0.989	0.534	0.458
17	1963_170 & 1979_158	2	0.512	0.492	0.951	0.648	0.993	0.793	0.744
18	1983_27 & 1983_37	2	0.640	0.527	0.960	0.809	0.995	0.784	0.774
19	1983_27 & 1979_33	2	0.672	0.581	0.957	0.685	0.994	0.739	0.725
20	1984_115 & 1981_49	2	0.520	0.500	0.963	0.788	0.992	0.733	0.598
21	1979_110 & 1989_233	3	0.368	0.351	0.935	0.551	0.991	0.648	0.663
22	1983_129 & 1976_176	5	0.428	0.266	0.954	0.703	0.992	0.651	0.573
23	1971_111 & 1972_291	5	0.445	0.393	0.931	0.566	0.993	0.656	0.513
24	1990_171 & 1988_88	5	0.275	0.297	0.893	0.602	0.993	0.658	0.558
25	1972_31 & 1984_115	5	0.533	0.536	0.947	0.754	0.991	0.694	0.581
26	1984_118 & 1971_336	5	0.479	0.356	0.960	0.681	0.991	0.808	0.609
27	1987_154 & 1964_144	5	0.501	0.492	0.954	0.846	0.991	0.665	0.527
28	1973_186 & 1986_218	5	0.392	0.393	0.926	0.586	0.989	0.645	0.498
29	1990_96 & 1990_171	5	0.325	0.439	0.932	0.724	0.992	0.689	0.732
30	1958_3 & 1992_144	5	0.399	0.372	0.909	0.551	0.992	0.664	0.476
31	1979_158 & 1965_111	7	0.586	0.529	0.964	0.755	0.994	0.670	0.606
32	1962_303 & 1972_291	7	0.394	0.540	0.931	0.672	0.988	0.745	0.613
33	1987_37 & 1989_233	7	0.169	0.234	0.903	0.560	0.992	0.565	0.530
34	1953_40 & 1953_24	7	0.867	0.836	0.989	0.931	0.996	0.763	0.700
35	1966_154 & 1976_43	7	0.434	0.431	0.947	0.745	0.989	0.588	0.663
36	1953_24 & 1957_52	7	0.259	0.177	0.883	0.357	0.985	0.437	0.418
37	1984_115 & 1971_49	7	0.489	0.482	0.942	0.817	0.993	0.714	0.662
38	1980_221 & 1984_115	8	0.489	0.539	0.944	0.727	0.988	0.726	0.615
39	1980_39 & 1969_324	8	0.663	0.648	0.973	0.933	0.992	0.652	0.575
40	1991_48 & 1987_189	9	0.517	0.537	0.943	0.858	0.993	0.635	0.634
41	1979_104 & 1979_110	9	0.793	0.695	0.974	0.922	0.994	0.831	0.802
42	1985_113 & 1969_324	9	0.690	0.619	0.972	0.941	0.981	0.561	0.518
43	1979_33 & 1979_110	9	0.815	0.838	0.990	0.949	0.995	0.776	0.799
44	1968_197 & 1972_62	10	0.425	0.584	0.914	0.687	0.993	0.806	0.764
45	1992_47 & 1984_115	10	0.518	0.540	0.945	0.725	0.993	0.733	0.688
46	1991_12 & 1985_113	10	0.755	0.725	0.980	0.952	0.990	0.615	0.601
47	1983_37 & 1979_33	10	0.754	0.750	0.978	0.884	0.993	0.679	0.721

Table 2
Pearson correlation coefficient for the word vector based and transformer based methods on Indian Supreme Court decisions (Gold standard pairs).

Sr. No.	Methods	Pearson Correlation Coefficient
Word Vector Based		
1	**TF-IDF**	0.614
2	Word2Vec	0.601
3	**Doc2Vec**	*0.685*
4	LDA	0.424
5	**fastText**	0.625
6	**GloVe**	0.567
Transformer Based		
7	**Longformer-base**	0.057
8	Sentence-BERT-large-nli	0.148
9	**Sentence-BERT-large-nli-stsb**	0.199
10	**Sentence-RoBERTa-large-nli-stsb**	*0.401*

References

[1] J. Huang, F. Gutierrez, H. J. Strachan, D. Dou, W. Huang, B. Smith, J. A. Blake, K. Eilbeck, D. A. Natale, Y. Lin, et al., Omnisearch: a semantic search system based on the ontology for microrna target (omit) for microrna-target gene interaction data, Journal of biomedical semantics 7 (2016) 1–17.

[2] H.-M. Müller, E. E. Kenny, P. W. Sternberg, M. Ashburner, Textpresso: an ontology-based information retrieval and extraction system for biological literature, PLoS biology 2 (2004) e309.

[3] P. D. Turney, Measuring semantic similarity by latent relational analysis, arXiv preprint cs/0508053 (2005).

[4] A. Schwering, Approaches to semantic similarity measurement for geo-spatial data: a survey, Transactions in GIS 12 (2008) 5–29.

[5] I. Matveeva, G. Levow, A. Farahat, C. Royer, Term representation with generalized latent semantic analysis, Amsterdam Studies in the Theory and History of Linguistic Science Series 4 292 (2007) 45.

[6] M. Oussalah, M. Mohamed, Knowledge-based sentence semantic similarity: algebraical properties, Progress in Artificial Intelligence 11 (2022) 43–63.

[7] S. Kumar, P. K. Reddy, V. B. Reddy, A. Singh, Similarity analysis of legal judgments, in: Proceedings of the fourth annual ACM Bangalore conference, 2011, pp. 1–4.

[8] S. Kumar, P. K. Reddy, V. B. Reddy, M. Suri, Finding similar legal judgements under common law system, in: International workshop on databases in networked information systems, Springer, 2013, pp. 103–116.

[9] S. Kumar, P. K. Reddy, V. B. Reddy, A. Singh, Similarity analysis of legal judgments, in:

Proceedings of the fourth annual ACM Bangalore conference, 2011, pp. 1–4.
[10] A. Mandal, R. Chaki, S. Saha, K. Ghosh, A. Pal, S. Ghosh, Measuring similarity among legal court case documents, in: Proceedings of the 10th annual ACM India compute conference, 2017, pp. 1–9.
[11] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).
[12] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: International conference on machine learning, PMLR, 2014, pp. 1188–1196.
[13] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.
[14] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the association for computational linguistics 5 (2017) 135–146.
[15] M. Ostendorff, E. Ash, T. Ruas, B. Gipp, J. Moreno-Schneider, G. Rehm, Evaluating document representations for content-based legal literature recommendations, in: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, 2021, pp. 109–118.
[16] L. Sahni, A. Sehgal, S. Kochar, F. Ahmad, T. Ahmad, A novel approach to find semantic similarity measure between words, in: 2014 2nd International Symposium on Computational and Business Intelligence, IEEE, 2014, pp. 89–92.
[17] Y.-S. Lin, J.-Y. Jiang, S.-J. Lee, A similarity measure for text classification and clustering, IEEE transactions on knowledge and data engineering 26 (2013) 1575–1590.
[18] S. Nourashrafeddin, E. Milios, D. V. Arnold, An ensemble approach for text document clustering using wikipedia concepts, in: Proceedings of the 2014 ACM symposium on Document engineering, 2014, pp. 107–116.
[19] M. Steinbach, G. Karypis, V. Kumar, A comparison of document clustering techniques (2000).
[20] R. Tous, J. Delgado, A vector space model for semantic similarity calculation and owl ontology alignment, in: International Conference on Database and Expert Systems Applications, Springer, 2006, pp. 307–316.
[21] P. D. Turney, Measuring semantic similarity by latent relational analysis, arXiv preprint cs/0508053 (2005).
[22] E. Giovannetti, S. Marchi, S. Montemagni, Combining statistical techniques and lexico-syntactic patterns for semantic relations extraction from text., in: SWAP, Citeseer, 2008.
[23] B. Perozzi, R. Al-Rfou, S. Skiena, Deepwalk: Online learning of social representations, in: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014, pp. 701–710.
[24] P. Qin, Z. Lu, Y. Yan, F. Wu, A new measure of word semantic similarity based on wordnet hierarchy and dag theory, in: 2009 International Conference on Web Information Systems and Mining, IEEE, 2009, pp. 181–185.
[25] M. S. Han, Semantic information retrieval based on wikipedia taxonomy, International Journal of Computer Applications Technology and Research 2 (2013) 77–80.
[26] G. Liu, R. Wang, J. Buckley, H. M. Zhou, A wordnet-based semantic similarity measure enhanced by internet-based knowledge., in: SEKE, 2011, pp. 175–178.