ISSN: 2633-4828 Vol. 5 No.4, December, 2023

International Journal of Applied Engineering & Technology

A COMPREHENSIVE SURVEY OF SEMANTIC TEXT SIMILARITY METHODS: HIGHLIGHTING THE NEED FOR A HINDI-ENGLISH TEXT SIMILARITY SYSTEM

Sharayu Mane[1], Durgesh Bhadane[2], Janhavi Jadhav[3], VandanaPabale[4] and Madhu Nashipudimath[5,e]

[1]manesharayu16@gmail.com, [2]durgeshbhadane1110@gmail.com,[3]janhavijadhav1425@gmail.com, [4]vandanapabale27@gmail.com and [5]madhu.mn@sigce.edu.in

[1]0009-0002-3353-4087, [2]0009-0001-2882-4576, [3]0009-0002-3622-7146, [4]0009-0000-9103-2638 and [5]0000-0001-8862-0826

ABSTRACT

With the exponential rise of textual data in the digital era, text similarity has become critical. This in-depth investigation investigates numerous strategies for assessing semantic text similarity andemphasizes the need for a specialized system to compare Hindi and English texts are available. This re- search examines the available literature, focusing on critical methods. A real-coded genetic algorithmtechnique for Hindi, similar to multi-model fine-grained nonlinear fusion Text summarization, as well as Word2Vec research in several languages. It also explores the application of LSTM and Siamese CNN models for predicting semantic textual similarity, in addition to examining the Cross-lingual plagiarism detection system. Despite past study, there are gaps, notably in Text similarity analysis between Hindi and English. This paper tries to fill these gaps by investigating the efficiency of the models (Word2Vec, LSTM, NMTScore, Bert) and paving the way for enhanced text similarity sys- tems with applications in various NLP domains. The foundation for developing precise text similaritysystems in the changing digital environment is laid out in this survey.

Keywords: Text similarity, Semantic similarity, Cross-lingual processing, LSTM, Word embeddings, NLP

INTRODUCTION

A key challenge in natural language processing (NLP) is assessing the similarity between texts based on their semantic similarities. It's crucial to investigate and develop specific systems for various lin- guistic systems in order to achieve accurate and efficient similarity computing. This thorough study sheds light on the difficulties and possibilities in this field by emphasizing the demand for a Hindi- English text similarity system. The development of semantic text similarity has greatly benefited fromadvances in deep learning techniques and large scale language models, including translation-based measures [3], genetic algorithms for text summarization [2], and word embedding techniques [19]. These developments also offer insightful information for examining the Hindi-English context. De- spite improvements in semantic text similarity techniques, extensive studies that focus on the demandfor a Hindi-English text similarity system are limited. Semantic text similarity in Hindi, which is widely spoken in India, has received little attention from previous studies since they have concen- trated on other languages, which has led to a lack of resources and study. This vacuum is filled by our survey, which focuses particularly on the Hindi-English setting and seeks to offer insights into current methodology, techniques, and difficulties in the computation of semantic text similarity.Through thisstudy, The aim is to examine the potential of current procedures, including NMTScore [3], Word2Vec [4], BERT [10], and LSTM-based approaches [9], and their suitability for the particular purpose of determining how similar Hindi- English texts are to one another. In order to understand cross- linguis-tic parallels and differences, also taking language- specific studies into account, as those on Bengali

[13] and Marathi [25]. The intention is to open the door for the creation of efficient and precise Hindi-

English text similarity systems. Numerous NLP domains, such as machine translation, text summa- rization, plagiarism detection, and information retrieval, can benefit significantly from the use of these systems.

RELATED WORK

The comparison of texts for similarity based on their meaning is a significant research area within Natural Language Processing (NLP). Over the years, researchers have developed a variety of method-ologies and models to

achieve the objective of determining semantic text similarity. This section thor- oughly reviews the existing literature, with an emphasis on the significant approaches that have beendeveloped.A model for computing semantic similarity utilizes the fusion of multiple fine-grained non- linear models [5].This method employs a number of models to collect semantic data from multiple perspectives, increasing the precision of similarity computation. A real-coded genetic algorithm wasused to solve an artificial Hindi text summarization problem[3]. By focusing on developing clear yet meaningful summaries, The application of semantic text similarity in summary tasks is provided[3].The semantic and morphologic similarities in embedded Ukrainian words are investigated using the Word2Vec technique [4]. This method demonstrates how semantic text similarity techniques can be used to a variety of languages. Although previous research and literature reviews have focused on semantic text similarity techniques, there is still a high demand for a Hindi-English text similarity system.The Siamese CNN and LSTM models are used to look into how to predict semantic textual similarity. A comparison between different deep learning architectures and how well they caught tex-tual semantic similarities is done [9]. The work teaches us how to use efficient models to measure textsimilarity. Furthermore, CrossLang a tool designed to identify plagiarism across languages, highlights the importance of cross-lingual text similarity in the context of interpreting linguistic similarities and differences[7]. Despite the extant literature, the topic of semantic text similarity still has research gapsand limitations. A key limitation is a lack of resources and dedicated research on the semantic text similarity of Hindi, a widely spoken language in India.This underscores the significance of conduct-ing thorough research to create a Hindi-English text similarity system. Moreover, even though somestudies have explored the applicability of models like Word2Vec [4] and LSTM [9], a comprehensive evaluation of their effectiveness in the realm of semantic text similarity remains necessary. This typeof analysis will provide useful insights into the strengths and limitations of these models, as well as drive future research into enhancing text similarity computation. By addressing these research gaps and constraints, a full study focusing on a Hindi-English text similarity system could develop NLP methods for determining semantic similarity. It will pave the way for the development of more effi- cient and accurate text similarity technologies. These techniques have broad applicability across var- ious domains in natural language processing, encompassing machine translation, text summarization, plagiarism detection, and information retrieval.

SEMANTIC TEXT SIMILARITY APPROACHES

To the extent of their semantic resemblance, two items, which can be concepts, sentences, or doc- uments, are said to be semantically identical to each other. It is an important linguistic strategy in Natural Language Processing (NLP). People form conclusions based on semantic similarities, which are similarities in semantics rather than form. Semantic similarity refers to how semantically similardistinct elements in a knowledge base are (such as linguistic units such as words, phrases, or con- cepts) [19]. A greater comprehension of the text can be reached when unstructured text is combined with lexical semantics, which is the study of word meanings. The initial phase in semantic analysis (i.e., dictionary definitions) is lexical semantics, which explores the meanings of specific words.The semantic analysis stage examines word relationships and determines the relevance of each sentence'ssections. Document classification, sentiment analysis, information retrieval, question-and-answer sys- tems, and plagiarism detection are some of the applications. Accuracy is a critical consideration in theprocess of detecting semantic similarity. A variety of methods can be used to assess the similarity oftext documents. Various strategies can be used to establish semantic-based similarity. These meth- ods use a variety of strategies to compare two texts semantically. The corpus-based technique uses

a statistical analysis of a huge corpus to assess how similar the words are. Deep learning algorithms can be used to determine the meaning of words by analysing a large corpus. A custom-made word semantic net underlies the knowledge-based strategy. This semantic net covers both word meanings and word associations. WordNet is a common semantic network used with apps. The third type of similarity is string-based. String similarity measures the degree to which a group of text strings and characters resemble one another. It calculates the distance between two text strings (similarity or dis- similarity)[9]. A variety of strategies can be used to calculate string similarity.This approach utilizesvarious techniques to achieve semantic similarity:

WORD2VEC

Word2Vec operates on the principle of distributional semantics, which posits that words found in sim- ilar contexts tend to share similar meanings. Word2Vec utilizes two primary architectures: Continuous Bag of Words (CBOW) and Skip-gram. CBOW predicts a target word based on the words in its sur- rounding context, while Skip-gram predicts the context words given a target word. To acquire these word embeddings, both architectures involve training a neural network on a substantial text corpus. Word2Vec is a powerful word embedding method that has significantly advanced natural language processing (NLP) by enabling the assessment of the semantic similarity between words. It represents the essential semantic connections among words as dense vectors in a continuous space.

The Word2Vec model has been extensively investigated for identifying semantic similarities among English words [19]. It explores various natural language processing techniques, including Word2Vec, for evaluating semantic similarity within context-specific domains. The study delves into the intrica- cies of Word2Vec and its potential applications in determining semantic similarity [19]. Word2Vec and Glove, with their dense word embedding representations, provide an effective means of capturingthe semantic nuances of words.

NMTSCORE

NMTScore is an interesting method for determining text similarity, especially when translation-based metrics are utilized. Using neural machine translation (NMT) models, this method computes similarity scores across texts published in several languages[3]. NMTScore incorporates both lexical and seman- tic parameters for determining translation quality and text alignment.A comprehensive multilingual evaluation of translation-based text similarity metrics, including the NMTScore [3].The study analy- ses how accurately the NMTScore evaluates semantic similarity between different language pairs. It throws insight on NMTScore's performance and usefulness by showing its benefits and drawbacks. The study investigates NMTScore's ability to detect plagiarism involving the distortion of original in-formation via translation. It emphasises NMTScore's potential as a useful tool for detecting instancesof hidden plagiarism.

BERT

BERT is a model for language representation.BERT is an acronym that stands for Bidirectional En- coder Representations from Transformers.The Bert model, developed by Google AI researchers, de- tailed how they employ a bidirectional LSTM layer combined with Masked LM (MLM) to capture the semantic and syntactic aspects of a text in a document.As a result, BERT can be utilised as an encoder to generate text embeddings that include semantic and syntactic features.The final score is de- termined by comparing the positional embeddings of each phrase to the Topics derived using LSA[8].To compute token similarity, BERT-score employs pre-trained BERT contextual embeddings.It evalu- ates the similarity by matching each token in the candidate sentence with every token in the reference sentence and calculating a similarity score. BERT embeddings have proven beneficial in numerous natural language processing (NLP) tasks. Rather than relying on precise string matching or heuristic

methods, the vector representation allows for a more flexible evaluation of similarity.BERT-score cal- culates the similarity of two phrases as the sum of cosine similarities between the embeddings of their tokens.[12] An examination of how well the BERTscore measures semantic similarity between differ-ent languages, as well as its advantages over other evaluation metrics, explains how the BERTscore works and demonstrates how valuable it is for measuring the quality of created texts[16]. Based on multivariate pattern recognition, this strategy recognises the significance of tools such as BERTscorein discovering and quantifying textual similarities.

LSTM

Long Short-Term Memory (LSTM) is a powerful RNN architecture designed to simulate sequential data. Because of its ability to capture long-range dependencies and keep information across longer se-quences, it is excellent for work involving semantic similarity analysis. For semantic textual similarity, a Siamese CNN (Convolutional Neural Network) and LSTM architecture-based predictive model is suggested[24]. The performance of many models is examined, as well as how well LSTM captures semantic commonalities across texts. Long Short-Term Memory

(LSTM) networks have showed sig-nificant potential in the domain of text similarity[25][26]. The innovative use of LSTM-based architec- tures to address the challenge of measuring similarity between short texts. With their ability to capture sequential relationships and contextual information, LSTMs are used as strong textual data encoding techniques. The LSTM encoder collects detailed semantic information from input texts, allowing thedetection of underlying linkages and similarities that conventional approaches could miss.The addi- tion of LSTM networks to the Siamese CNN-LSTM hybrid model improves understanding of semantic similarity and results in more accurate predictions. LSTM-based algorithms have proven to be highly effective in enhancing the computation of text similarity. This progress opens up new possibilities for improved applications across various natural language processing tasks, such as recommendation systems, information retrieval, and content summarization.The reason for the effectiveness of LSTM lies in the fact that the true meaning of a phrase in all natural languages is influenced, in part, by the specific order in which words are used. LSTM excels at comprehending sentences by incorporating the sequential information of each word.In this model, a fully connected neural network serves as the output layer, which prompts the output layer to adjust its parameters to accurately learn patterns from the features generated by LSTMs. This is a departure from the baseline method, where cosine similarity is used as the final layer.

The bidirectional structure also plays a crucial role in encoding the features of each input text

more comprehensively. Each LSTM instance utilized in our proposed model operates independentlyand maintains a unique set of parameters [26].

DISCUSSION

Multilingual text similarity can be extremely helpful in activities like machine translation and infor- mation retrieval. Comparing texts in different languages is a contemporary challenge. It is mentionedas a vital stage in the preprocessing of text content to apply natural language processing techniques. The extraction of relevant information from the texts depends heavily on techniques like filtration, stop words removal, lemmatization, part of speech tagging, and named entity identification [9] [4].Most of the existing approaches were based on linguistic pattern, syntax based features, and translation techniques. Predominantly all the state-of-the-art methods are based on deep learning models such word2vec for representation, convolutional neural network, recurrent neural network, long short term memory (LSTM) etc.The benefits of using Word2Vec for cross-lingual plagiarism detection includeimproved accuracy and the ability to detect plagiarism even when the plagiarized document is in a different language[21]The maximum sequence length supported by NMT models for NMTSCORE is often relatively short.NMTSCORE is slower compared to baseline measures[3].Multi-encoder NMTcan be used in text similarity by using multiple encoders to decode a sentence. Each encoder embeds

the meaning of the sentence as a vector and sends it to a decoder, which processes the vectors and emits the translation. This approach can improve the quality of translation by utilizing current infor-mation such as word frequency and part-of-speech. However, specific details on how multi-encoder NMT can be specifically used for text similarity are not mentioned in the search results.[24]For lstm on Quora Dataset LSTM achieved an accuracy of 87.5% and F1 score of 87.4%.Both the accuracy and F1 score achieved by the LSTM on this dataset are higher than the corresponding metrics achieved by the baseline methods. This indicates that the proposed algorithm is more effective at making accurate pre- dictions and achieving a balance between precision and recall compared to the baseline methods on the Quora dataset.For lstm MSR Paraphrase Corpus Dataset LSTM achieved an accuracy of 80.4% and anF1 score of[26].The development of a bidirectional Long Short-Term Memory (LSTM) based method to calculate cross-lingual semantic text similarity for long text and short text shows that bi-LSTM based obtains better performance on short text data such as news titles and alert messages, which areon average shorter than 20 words, in contrast to normal news articles with more than 200 words on av-erage.[26]BERTLARGE is another pre-trained model that is trained on the same data as BERTBASEbut with a larger model size, including more layers, hidden size, and attention heads.BERTBASE is a pre-trained model that is trained on the BooksCorpus (800M words) and Wikipedia (2,500M words) with a batch size of 128,000 words. It learns [SEP], [CLS], and sentence A/B embeddings during pre- training.BERTBASE and BERTLARGE outperform all systems on all tasks by a

substantial margin,obtaining 4.5% and 7.0% respective average accuracy improvement over the prior state of the art.

CONCEPTUAL SOLUTION

The Fig1 aims to address the challenge of linguistic differences by implementing a thoughtful trans-lation process and leveraging advanced semantic algorithms to achieve more precise text similarity assessments.

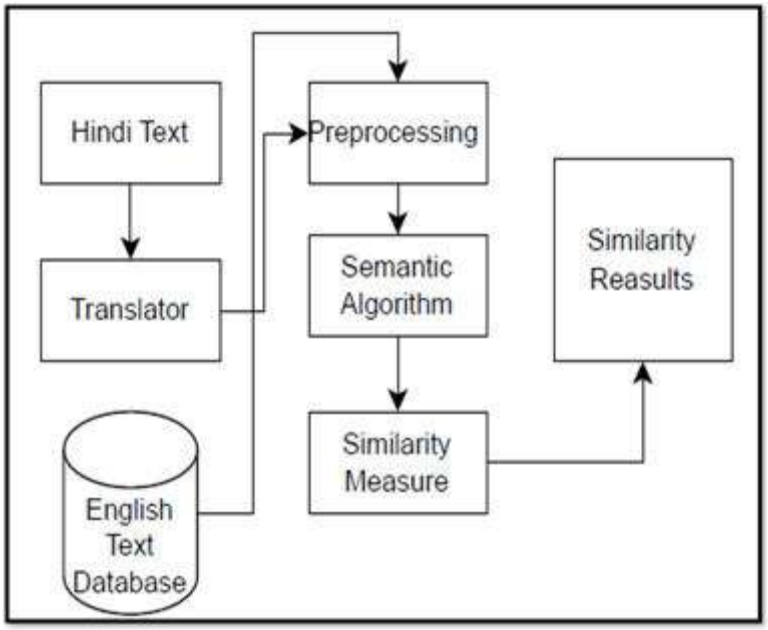

Fig. 1: CONCEPTUAL SOLUTION.

- **Data Input (Hindi documents):** These documents may contain diverse linguistic characteris- tics, dialects, and cultural nuances.
- **Translation (Hindi documents into English):** To enable cross-lingual comparisons, Hindi doc- uments are translated into English for common language analysis.
- **Preprocessing:** This stage involves cleaning and preparing the translated documents.
- **Semantic Algorithm:** The core of the framework is the semantic algorithm, which uses tech- niques like Word2Vec, BERT or LSTM to assess semantic similarities between translated texts.
- **Similarity Measures:** This is achieved using methods like cosine similarity, Jaccard similarity,or other distance metrics, aiming for a clear and interpretable similarity score.
- **Similarity Result:** The final output provides a similarity score or metric that signifies how alike the two texts are, despite the linguistic disparities.

CONCLUSION

In this comprehensive survey of semantic text similarity methods, the importance of multilingual textsimilarity is highlighted, given its applications in areas such as machine translation and information retrieval. The comparison of texts across languages is a significant challenge that affects various nat- ural language processing tasks. Techniques like linguistic pattern analysis, syntax-based features, andtranslation methods have traditionally been employed. However, modern approaches predominantly rely on deep learning models, including Word2Vec for representation, convolutional neural networks, recurrent neural networks, and LSTM.Word2Vec and LSTM are particularly well-suited for measuring text similarity in Hindi-English due to their unique strengths. Word2Vec's

ability to capture seman- tic relationships between words and represent them as dense vectors transcends language barriers, enabling effective cross-lingual comparisons. This is crucial for languages with different linguistic structures like Hindi and English. On the other hand, LSTM's sequential nature makes it adept at cap- turing contextual dependencies within sentences, a critical aspect in assessing text similarity. Given the syntactic and grammatical differences between the two languages, LSTM's capacity to capture intricate patterns and contextual nuances enhances its applicability. Together, Word2Vec and LSTMoffer a synergistic approach, combining cross-lingual semantic representation and sequential contextmodeling to navigate the challenges posed by Hindi-English text similarity, ultimately yielding moreaccurate and meaningful results than other models.

REFERENCES

[1] Jain, A., Arora, A., Morato, J., Yadav, D., Kumar, Automatic text summarizationfor Hindi using real coded genetic algorithm. Applied Sciences, 12(13) (2022), 6584.

[2] Muneer, I., Iqra, Nawab, Cross-Lingual Text Reuse Detection at sentence levelfor English–Urdu language pair. Computer Speech and Language, 75 (2022), 101381.

[3] Vamvas, J., Sennrich, R., NMTScore: A Multilingual Analysis of Translation-based Text Similarity Measures, 2204 (2022), 13692.

[4] Savytska, L. V., Vnukova, N. M., Bezugla, I. V., Pyvovarov, V., Subay, Using Word2vec technique to determine semantic and morphologic similarity in embedded words of the Ukrainian language, 2021

[5] Agarwala, S., Anagawadi, A., Guddeti, R., Detecting Semantic Similarity Of Doc uments Using Natural Language Processing. Procedia Computer Science, 189 (2021), 128-135.

[6] Zhang, P., Huang, X., Wang, Y., Jiang, C., He, S., Wang, Semantic similarity com-puting model based on multi-model fine-grained nonlinear fusion. IEEE Access, 9 (2021), 8433-8443.

[7] Gupta, H., Patel, Method Of Text Summarization Using Lsa And Sentence Based Topic Modelling With Bert. ICAIS 2021, pp. 511-517, IEEE.

[8] Tiyajamorn, N., Kajiwara, T., Arase, Y., Onizuka, Language-agnostic Representation from Multilingual Sentence Encoders for Cross-lingual Similarity Estimation. 2021, pp.7764-7774.

[9] Wang, X., Dong, X., Chen, S., Text duplicated-checking algorithm implementation based on natural language semantic analysis. ITOEC 2020, pp. 732-735, IEEE.

[10] Roostaee, M., Sadreddini, M. H., Fakhrahmad, An effective approach to candidate retrieval for cross-language plagiarism detection: A fusion of conceptual and keyword-based schemes. Information Processing and Management, 57(2) (2020), 102150.

[11] P, Sunilkumar Shaji, Athira., A Survey on Semantic Similarity. 1-8. 10.1109/ICAC347590.2019.9036843, 2019.

[12] Devlin, J., Chang, M. W., Lee, K., Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[13] Koroteev,, On the Usage of Semantic Text-Similarity Metrics for NaturalLanguage Processing in Russian. 2020, pp. 1-4. IEEE.

[14] Nair, Nair, Nair, Prabhu, Kulkarni, semantic plagiarism detection system for english texts. IRJET 2020.

[15] Xia, C., He, T., Li, W., Qin, Z., Zou, Z., Similarity analysis of law documents based on Word2vec. 2019, pp. 354-357, IEEE.

[16] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., Artzi, Bertscore: Evaluating textgeneration with bert. 2019 pp 1904.09675.

[17] Haneef, I., Adeel Nawab, R. M., Munir, E. U., Bajwa, Design and development of a large cross-lingual plagiarism corpus for Urdu-English language pair, 2019.

[18] Bakhteev, O., Ogaltsov, A., Khazov, A., Safin, K. and Kuznetsova, R.,"CrossLang: the system of cross-lingual plagiarism detection." 2019.

[19] Jatnika, D., Bijaksana, M. A., Suryani, Word2vec model analysis for semantic similarities in English words. 157 (2019), 160-167.

[20] Som, S., Analysis of Natural Language Processing (NLP) approaches to determine se- mantic similarity between texts in domain-specific context (Master's thesis), 2019

[21] Agarwal, Cross-lingual plagiarism detection techniques for English-Hindi language pairs. Journal of Discrete Mathematical Sciences and Cryptography, 22(4) (2019), 679-686.

[22] S¸enel, L. K., Utlu, I., Yucesoy, V., Koc¸, A., C¸ukur, Generating semantic similarity atlas for natural languages, 2018 pp. 795-799. IEEE.

[23] Xylogiannopoulos, K., Karampelas, P., Alhajj, Text mining for plagiarism detection: multivariate pattern detection for recognition of text similarities. 2018, pp. 938-945, IEEE.

[24] Yeong, Y. L., Tan, T. P., Gan, K. H., Mohammad, Hybrid machine translation with multi-source encoder-decoder long short-term memory in english-malay translation. (2018) 8(4-2), 1446-1452.

[25] Pontes, E. L., Huet, S., Linhares, A. C., Torres-Moreno, Predicting the semantic textual similarity with siamese CNN and LSTM. (2018) pp.1810.10641.

[26] Yao, L., Pan, Z., Ning, H., Unlabeled short text similarity with LSTM encoder. (2018) 7, 3430-3437.