An Artificial Intelligence Approach for Word Semantic Similarity Measure of Hindi Language

Farah Younas[1], Jumana Nadir[2], Muhammad Usman[3], Muhammad Attique Khan[4], Sajid Ali Khan[5], Seifedine Kadry[6], Yunyoung Nam[7]
[1,3]Department of Computer Science, Shaheed Zulfikar Ali Bhutto Institute of Science and Technology, Islamabad 44000, Pakistan
[e-mail: farah.younas@szabist-isb.edu.pk; dr.usman@szabist-isb.edu.pk]
[2]Computer Engineering Department, San Jose State University, San Jose, CA, USA
[e-mail: jumanazoeb.nadir@sjsu.edu]
[6]Department of Computer Science, HITEC University Taxila, Pakistan
[e-mail: attique.khan@hitecuni.edu.pk]
[5]Department of Software Engineering, Foundation University, Pakistan
[e-mail: sajid.ali@fui.edu.pk]
[6]Faculty of Applied Computing and Technology, Noroff University College, Kristiansand, Norway
[e-mail: skadry@gmail.com]
[7]Department of Computer Science and Engineering, Soonchunhyang University, Asan, Republic of Korea
[e-mail: ynam@sch.ac.kr]
*Corresponding author: Yunyoung Nam

Received September 12, 2020; revised November 9, 2020; revised December 21, 2020; accepted February 11, 2021; published June 30, 2021

Abstract

AI combined with NLP techniques has promoted the use of Virtual Assistants and have made people rely on them for many diverse uses. Conversational Agents are the most promising technique that assists computer users through their operation. An important challenge in developing Conversational Agents globally is transferring the groundbreaking expertise obtained in English to other languages. AI is making it possible to transfer this learning. There is a dire need to develop systems that understand secular languages. One such difficult language is Hindi, which is the fourth most spoken language in the world. Semantic similarity is an important part of Natural Language Processing, which involves applications such as ontology learning and information extraction, for developing conversational agents. Most of the research is concentrated on English and other European languages. This paper presents a Corpus-based word semantic similarity measure for Hindi. An experiment involving the translation of the English benchmark dataset to Hindi is performed, investigating the incorporation of the corpus, with human and machine similarity ratings. A significant correlation to the human intuition and the algorithm ratings has been calculated for analyzing the accuracy of the proposed similarity measures. The method can be adapted in various applications of word semantic similarity or module for any other language.

This work was supported by the Soonchunhyang University Research Fund.

http://doi.org/10.3837/tiis.2021.06.006 **ISSN : 1976-7277**

Keywords: Artificial Intelligence, word similarity; semantic nets; natural language processing; corpus; synonymy

1. Introduction

Hindi is the world's fourth most widely spoken language and has more than 295 million native speakers. Hindi is the official language of the Indian subcontinent and is also widely spoken in countries like Mauritius, Nepal, and Fiji. Hindi script is written in Devanagari and uses more Sanskrit words. The Devanagari script is used for over 120 Indian languages. Most of the authoritative and legal scripting in India is in Hindi, along with the associate language- English. The Hindi word 'avatar' meaning incarnation or taking a new form is being used in Computer Sciences and Artificial Intelligence (AI) for robotics [1]. According to the findings of the Indian Human Development Survey, 82% of adult females and 72% of adult males in India did not speak English [2]. This survey demonstrates the need to create more intelligent interactive models that guide the non-English speakers through the smooth operation of computer applications.

The prerequisite to developing Conversational Agents in Hindi is the thorough understanding of the Language and review of its available resources; and the need to develop computational models for calculating semantic similarity between Hindi concepts. In the field of Natural Language Processing, determining the semantic similarity between any two words has always been enigmatic. Breakthrough approaches for lexical ontology have been introduced for the English Language. However, there is no benchmark approach for the Hindi language.

The term Semantic Similarity or Semantic Relatedness in precise is a metric defined to accentuate the likeness or closeness to predict the vulnerable similarity between a pair of words. The semantic similarity has gained a widespread application in the area of Natural Language Processing for its usage in level segmentation, documentation clustering, text summarization, and annotation [3]. NLP uses the common techniques of Cosine Similarity of TF-IDF between the documents or words, but effective methods that compute semantic relatedness between sentences or words are required.

The application areas of Natural Language Processing require effective methods that compute semantic relatedness between sentences or very short texts. Information retrieval from the internet and dialogue systems are such major applications [4, 5]. This paper contributes to the field of Short Text Semantic Similarity for Hindi Language, by proposing a similarity measure that calculates semantic relatedness between words. Firstly, a detailed review of the available resources for Hindi has been done. The largest available Hindi Corpus HindMonoCorp is used in the experimental methodology as a semantic knowledge base, which contains approximately 44 million Hindi sentences.

Groundbreaking advances and research have been performed in AI related to text [6, 7]. The English Language, including Dialog systems and Conversational Agents [8, 9] are the major aspects. But when it comes to other languages, one argument put forward is that; Can the expertise obtained in the English Language be transferred to other resource-poor languages?

A spearheading investigation has been done for the English Language by H. Rubenstein and J. Goodenough, that 'the more similar the words are in their meaning, the more similar they are in their contextual distributions' [10]. Their research centers on comparing the contextual distributions of a word pair. This data has been trusted and used in intervening years because of its stability. In relation to Hindi, the presumptive questions are:

- Whether Hindi can accept the challenges of developing Conversational Agents?
- Whether a word semantic similarity measure can be obtained for Hindi with the available resources?

- *H_0 (null hypothesis):* A dissent that a word semantic similarity measure cannot be built for Hindi.
- *H_1 (alternative hypothesis):* An assertion that adequate lexical resources are available for Hindi. And the experimental evidence suggests that the null hypothesis is false.

Since very limited lexical resources are available for resource-poor Hindi language and not much work is done in Natural Language Processing, there is a need to have a more concrete semantic similarity measure. Moreover, there is no AI conversational agents available for Hindi so a robust conversational AI agent should be developed to support the Hindi language. If the proposed measure provides promising result and prove to be successful for the Hindi language, similar advances can be made for several other Indian languages which are derived from Hindi language e.g., Tamil, Marathi, etc.

There are very few lexical or semantic similarity works in literature focusing on the Hindi language as it is a resource-constrained language, an important reason being the unavailability of large, structured data or corpus. For developing Conversational Agents in Hindi, a thorough understanding of the language and a review of its available resources is important. Nowadays semantic similarity between words can be found using many pre-existing libraries, but a benchmark can be achieved when we involve its native speakers. Therefore, in this literature, we prepare a benchmark approach that adapts the English benchmark as a stepping-stone, validating it by its native speakers.
The generalizability of NLP models to different languages can be achieved in various scenarios. Many models are said to perform well for the English language, however, there are not many that generalize well for other languages such as Hindi. Building high computational models with such noisy data of such low-resource languages tends to give poor results. Therefore, an adaptation of a good semantic similarity paired dataset from a benchmark English approach can be useful and groundbreaking towards solving this problem. Semantic similarity has gained widespread applications in NLP for its usage in level segmentation, documentation clustering, text summarization, and annotation. We believe that our proposed approach can help in achieving state-of-the-art results in building these applications for the Hindi language.

Thus, following to this need, a word semantic similarity measure is proposed in this work. And an attempt to support the H_1 hypothesis is carried out and tested in the experimental analysis. This research introduced a similarity measure to calculate the degree of contextual overlap between word pairs. It also introduced a lexical ontology approach for Hindi Language by carrying out the comparison of the largest available Hindi corpus and Human judgments of degree of similarity present in benchmark word pairs. It is benchmark approach for Hindi adapted from well-known English languages approach. This study developed a more intelligent and interactive model to guide speakers who do not speak English language to perform smooth computer application operations. One of the major contributions of this study is the accurate translation of benchmark dataset that can be adapted by various NLP based applications.

Table 1. Notations

Notation	Explanation
E_x	English word pairs
H_x	Word pairs translated into Hindi
A_x	Set of words containing A
B_x	Set of words containing B
W_i	Input text
V_i	Weight vector
K_i	Inverted index

The remaining sections of this paper are organized as follows. Section 2 is Related work. Section 3 presents the Proposed Methodology. Section 4 elaborates the results and its evaluation. Finally, the conclusion is mentioned in Section 5.

2. Related Work

There exists an extensive literature on measuring short text similarity for English [11-13], whereas lesser work has been done for Hindi in this domain. As this thesis is based upon using Corpus as a knowledge base, this section demonstrates the related literature that uses different resources. The advantages and limitations of each method are also discussed. In this way, a comparison of the aforementioned methods is obtained. This section also inspects and reviews some work relating to the evaluation of short text semantic similarity in Hindi, to probe the strength and weakness in evaluating short text similarity. This section is therefore categorized into three sub sections-a brief definition of NLP and Semantic Similarity, different linguistic approaches with different knowledge resources and other descriptive literature work presented for the Hindi Language.

2.1. Natural Language Processing (NLP)

Natural Language Processing or simply abbreviated as NLP is a branch of computer science that deals with the mutual understanding of the interaction between the computer and human spoken language. It concenters the whole essence of complying with the computer to be able to communicate with people using everyday language. NLP is an elemental part of Artificial Intelligence (AI), also called Computational Linguistics that helps in concatenating the computational methods to aid in local human language understandability. This amends to the revolution in the field of computational science where a machine could actually respond to human language. In other words, NLP is the ability of any computer machine to understand human language as an input and derive sensible meaning from it. SHRDLU and ELIZA are a few of the successfully developed NLP systems in the 90s [14, 15].

Today's NLP algorithms use statistical machine learning where many different classes of machine learning algorithms are applied. The goal of the procession would require the human to communicate in the form of voice and machine on the other hand translating the information into strings of words in the desired natural language. It requires the human to speak in a manner such that the language is unambiguous and highly precise and structured in terms of programming. However, it is a tough task since human speech is not always precise and is often ambiguous because of slang and different regional dialects. Thus, ambiguity in the language plays an important role in the computational language and sometimes becomes erroneous or unpredictable. The current approach in the dilution of machine learning is a type of artificial intelligence that uses the data patterns to improve the program understanding. A program goes through a series of operations under software tasks. To name a few, when any sentence is given an input; it is firstly segmented, processed as a part-of-speech tagging, and parsed sequentially. It involves the identification of the named entities, recognizing the data, and mapping it in proper order.
The research analysis of the NLP is broadly correlated to the study of Artificial Intelligence (AI). Though the gradual shifting to compositional semantics from lexical semantics of NLP research is corresponded, however, the human-level natural language processing is a complete artificial intelligence conundrum. It is similar to correcting the intricateness of artificial intelligence delusions for the computers to act as an intelligent person. Hence it is agreed to the fact that as the proficiency in artificial intelligence surges it would correspondingly augment the development of NLP's future [16].

2.2. Semantic Similarity

Semantic similarity also called semantic relatedness, is the degree of measure of how scrupulously two words or sentences cognate with each other. To assay the related closeness and estimate the semantic similarity between the two terms, it is often construed by defining topological similarities using the methodology of ontology. In other words, semantic similarity is the estimation of semantic distance between two concepts based upon the predefined ontology. For example, the words apple and orange are more related to each other because of being hyponyms than the words apple and table. Several factors and calculations are studied in order to measure the semantic similarity [17].
Node-based and Edge-based are a few of the many measures for calculating semantic similarity. Most of these measures are evaluated using the reliable lexical ontology WordNets. Besides this, the Corpus-based (or information theory-based) and the Wikipedia based measures are also extensively practiced. These methods are proven useful in specific applications of linguistic studies. The literature works that used these methods are briefed in this section [18, 19].

2.2.1. *Corpus based Semantic Similarity approach*

This method of corpus-based semantic similarity is based on the fact that it strives to identify the degree of similarity between the words by extracting the information from the large corpora exclusively. Usually in this approach, the word relationships are often procured from their simultaneous occurrence in the corpus. The corpus-based semantic focuses on the exemplification of linguistic expressions in terms of distributional properties. This method is aimed at large because of its vast coverage. Pointwise Mutual Information (PMI) and Latent

Semantic Analysis (LSA) are generally two types of metrics used under this similarity method [20, 21].
The PMI approach is based on sorting out the neighboring words out of the two target words. It was first used by Church and Hanks to measure the word similarity. This method uses AltaVista advanced search query syntax to formulate the probabilities. The words that are common in the list are distinguished and the respective PMI values are approximated to calculate the corresponding similarity between the words. It is one of the simplified methods of computing the similarity of words for a corpus-based approach. A formula is proposed in their work that calculates the statistical dependence between the words.
The large corpora are broken down into small chunks of texts roughly to the size of the paragraph where these contexts are analyzed and the frequency of occurrence of each word is computed in the form of matrix format with a column for each word and row for each paragraph. The theorem of SVD is applied to X (Where X is a cell or word by context matrix) using the following formula:

$$X = W\ S\ P\ I \tag{1}$$

Where W and P are in column orthogonal form and S is the diagonal matrix for non-zero entries. On a further note, the whole matrix is compressed to a dimensional space of matrices. The degree of similarity between two words is therefore measured using LSA by taking the cosine of the angle between the corresponding row vectors. Thus, LSA is sometimes also considered as an alternative method to overcome hitches in the standard vector space model.
The main advantage of corpus-based semantic similarity method is it does not require any handmade resources. Apart from having a large and apt corpus it does not prompt for any issues regarding the completeness of the resources. According to them, in general, the corpus-based measure gives better performance in recall as compared to other semantic similarity-based measures.

2.2.2. *Wikipedia based Semantic Similarity approach*

The computation of semantic similarity among texts requires the assessment of a huge amount of world knowledge. Thus, Wikipedia based semantic similarity is one such method proposed which is based on the prospects of Explicit Semantic Analysis (ESA), the novel method serves the purpose of texts meaning in its contextual manner in a high-dimensional space [13]. This method explicitly expresses the context of any text as a weighted vector of Wikipedia based concepts using machine learning techniques. The Wikipedia based semantic approach found its primary application in processing natural language texts that incur to the bulk amount of data to compute relatedness in relevance. Wikipedia accounting to one of the largest information databases in existence help in the representation of concepts in a higher-dimensional space which involve various classification techniques that culminate to represent the forehand meaning of the text in the terms of Wikipedia based concepts [14].

A Semantic interpreter is constructed using machine learning techniques to form the weighted sequence of Wikipedia concepts by mapping the fragmentation of texts. The input texts are given in the form of plain text. This helps in availing the method of conventional classification to order the articles according to their relevance with the respective text fragment which permits us to use the encyclopedia directly without requiring any deep understanding. Every Wikipedia subject is equated as an attribute vector of words whose weights are assigned using Term Frequency- Inverse Document Frequency (TFIDF) scheme. To eliminate the inappropriate relationship between the words, an inverted index is built up which plays an epic role in mapping every word that appears in a list of concepts. This methodology also helps in speeding up the semantic interpretation level [3]. Consider the following abbreviations to manipulate the semantic interpretation:

Let $T = w_i$ input text

$< V_i >$ is weight vector

$< K_j >$ be inverted index

$C_j \left\{ c_j \in c_{1,\dots} c_N \right\}$ be total number of Wikipedia concepts.

Thus, the semantic similarity interpretation of Vector V for text T is defined by:

$$\sum w_i \in T\ V_i . K_j \tag{2}$$

The semantic similarity is computed by comparing the vectors of text fragments taking the cosine metric. Humans do have a congenital tendency to judge semantic relations between texts. However, to further contemplate on the Wikipedia based approach; another semantic interpreter based on a large knowledge repository was introduced that is referred to as Open Directory Project or ODP. ODP is one of the largest web directories to date including textual description and URLs accounting for 436 Mb of text.

The experimental-based tests show that Wikipedia-based semantic interpretation surpasses the ODP-based and is superior to it. The main objective that the Wikipedia-based approach has used widely is its largest knowledge-based repository on the web. A study also claims Wikipedia is superior to Britannica in terms of accuracy. Most importantly, Wikipedia is available in a variety of languages, English being the most predominant. Also, it is comparatively easy to interpret the model using the Wikipedia-based method.

2.3. Natural Language Processing and Semantics in Hindi Language

This section enlightens the related work contributed to the field of Hindi semantic similarity and Natural Language Processing as a whole. In a work by J Singh and A. Sharan, a computational model has been proposed, based upon lexical ontology for the Hindi language [22]. The proposed model incorporates Hindi WordNet module and uses Miller and Charles's benchmark dataset. The work mainly focuses on the similarity aspects of ontologies such as WordNet. A Lexical Ontology and Semantic Similarity framework are projected. This framework uses the WordNet taxonomy to produce a hypernym tree finding module. Eventually, the work centers around comparing the Edge-based and Node-based semantic similarity measures by using the Wordnet knowledge resource.

Their work suggests that the Indo WordNet (Hindi Wordnet) quantifies the semantic networking systems specifically for Indian languages. Usually, every individual language designs its own WordNet structure however they are interlinked with interlingual links and is directed to Interlingual Index (ILI). The lexical information derived from the Hindi WordNet in respect of the word's meanings is termed as lexicons. The tree-level specification of the Hindi WordNet hypernym concept is diagrammatically represented as in **Fig. 1**.

The work follows a set of experiments for comparing the Human similarity measures and similarity algorithms. As observed from the experimental result, a lot of variation in the result is foreseen and a lot of rectification is demanded for the Hindi text.

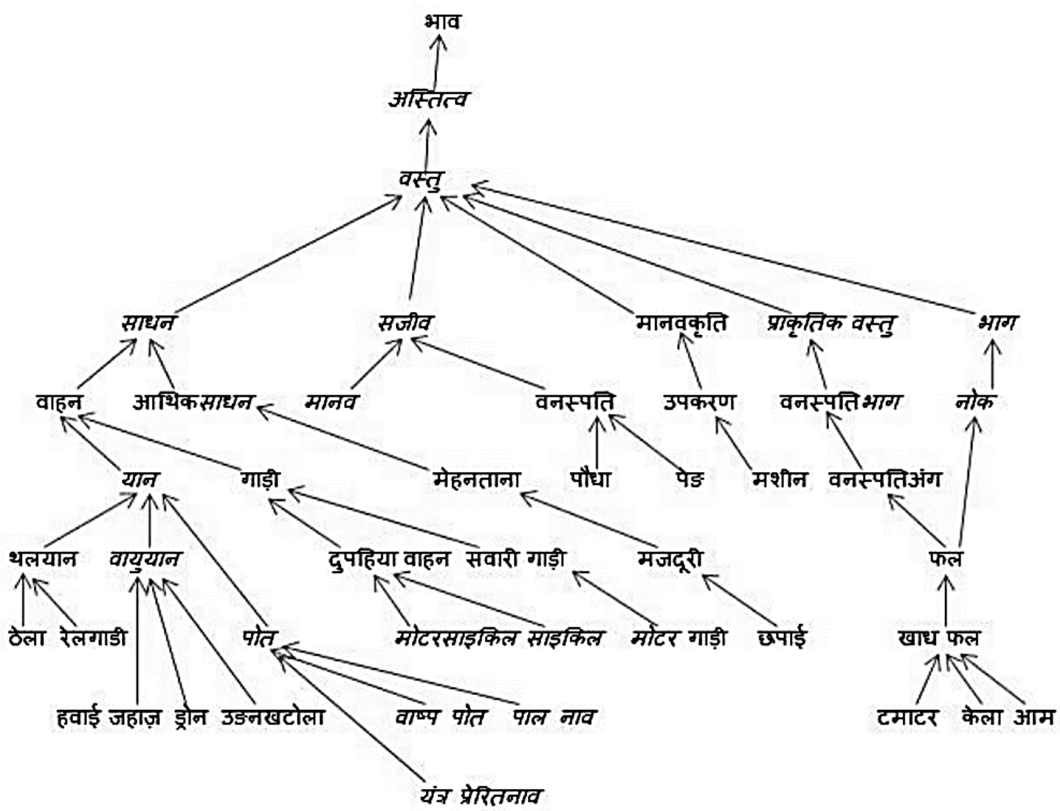

Fig. 1. Hindi WordNet Hypernym Tree

The literature review shows a big scope for the Hindi Language and the need to produce more effective semantic similarity measures for sentences, very short texts and words. The next section performs a thorough investigation of lexical resources available for Hindi.

2.4. Review of Resources

Wordnets and Corpora are said to be the supreme resource or knowledge banks of a Language. They are an essential part of the computations involved in the mainstream Information Extraction, Natural Language Processing, and Word Sense Disambiguation, etc. WordNets are composed of lexical structures whereas, a Corpus or Corpora (plural) on the other hand consists of 'real-world' instances. A survey shows that extensive resources are available for Hindi and also other Indian Languages [23]. WordNet and Corpus knowledge bases for Hindi are thoroughly investigated in this section.

2.4.1. *Wordnet*

WordNet is a combination of Synsets i.e., synonyms that are linked by semantic relations. Adjectives, nouns, verbs, and adverbs are assembled into sets of cerebral Synsets. Wordnet built for English at Princeton University was the first. It was developed over the years with a view of machine translation. Then it was followed by wordnets for other European Languages. The influence of the English WordNet has inspired the development of wordnets for many other languages in the world; one such language is Hindi [24].

2.4.2. *Hindi WordNet*

The linguistic research in this paper manifests the existence of vast resources available for Hindi and other Indian Languages. With the initiative of the Hindi WordNet by Pushpak, a vast majority of wordnets were started to build for other Indians. Languages. The design of the Hindi WordNet was inspired by the English WordNet. It was developed by the Centre for Indian Language Technology (CFILT) at the Indian Institute of Technology Bombay. Splendid features such as antonymy and meronymy are introduced in the dataset. Synonyms are linked through prominent semantic relationships. Also, a conducive underlying database design was formulated. The part-whole relationship is devised in eight different ways. Keeping in mind the varied nature of the Hindi Verbs, the entire verb lexicon is structured. As of March 2010, the Synsets Linkage and the number of unique words covered by the Hindi WordNet was 33900/82000. The Wordnet is in constant headway by increasing the Linked Synsets and Bilingual Mappings. At present, the Wordnet has 25863 Linked Synsets and 39036 total Synsets [25, 26].

2.4.3. *Indo WordNet*

Indo WordNet is an enhancement of the Hindi WordNet by adding wordnets of 17 Indian Languages. Indo WordNet contains languages such as Assamese, Bengali, Bodo, Nepali, Manipuri, Punjabi, etc. These languages are spoken by millions of people in India and its neighboring countries like Bengali in Bangladesh and Nepali in Nepal etc. The merge and expansion approaches are adapted in building Indo WordNet. The wordnet mainly follows the design principle of the English Wordnet. Special attention is paid to language-specific phenomenon such as complex predicates. The Hindi and Marathi language Synsets are created upon the ethics of Coverage, Minimality, and Replaceability. 16 out of 22 official Indian Languages have started making wordnets using the expansion approach, namely Punjabi WordNet, Marathi, Sanskrit, Oriya, Dravidian, Gujarati, and North East WordNets etc. **Table 1** shows the core and decisive Synset linkage and number of unique words covered in the Wordnet by 1st March 2010 [27, 28].

2.4.4. *Hindi Corpus*

The Technology Development for Indian Languages (TDIL) Group has presented a Corpora, which has a corpus collection of 10 Indian Languages including Hindi. The Hindi Corpus has an approximate of 3 million words. The count of distinct words that is words without repetition is 127241. The average word length is 4.69 approximately. The corpus also describes the cumulative frequency of the 4 most repeated words in the corpus, such as prepositions [29].

The representation of the corpus includes headers containing the authors, source, and publishers, etc. The corpus is also divided into subject-wise categories such as Social Sciences, Literature, Mass Media, Fine Arts, and Novel. Many texts consist of randomly chosen publications during the year 1980-1990. The corpus is not freely available for download on the website but can be obtained by contacting the TDIL Group, Ministry of Information Technology [4].

3. Proposed Methodology

3.1 Architecture

Fig. 2 gives the outlook of the work executed in this paper. The proposed method derives semantic information by comparing the texts from a large database. The sequence of words that carries useful information is known as the text.

After the scrutiny of the knowledge base for Hindi, the candidate contexts from the knowledge bank are used in the context's extraction schema. The contexts extraction module works by selecting the sentences from the corpus, based on a specifically given criterion. These sentences are augmented into the proposed overlap measure, and then the machine overlap measure is calculated. The machine overlap measure is calculated by building a Python code that handles the Unicode texts and a large database. Detailed working of the code is also explained in the coming section. The results of the proposed algorithm are then compared with the recorded Human similarity scores. Finally, a qualitative assessment is done using the correlation coefficients to rectify the results.

The following sections present a detailed implementation of each of the above-mentioned steps. A detailed procedure of building a dataset is explained in the next section, followed by the set of experiments performed.

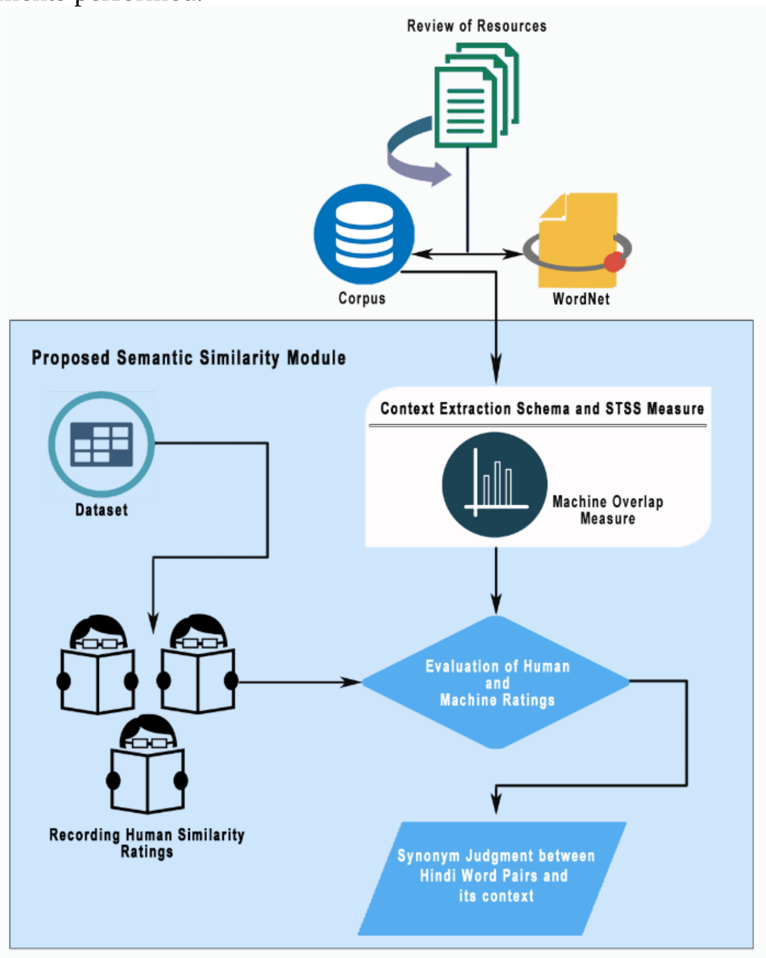

Fig. 2. Proposed Architecture

3.2. Proposed similarity measure

The research is concerned with finding relationships between a word pair and its contextual associates, as an approach for short text semantic similarity in Hindi. In other words, an estimation using statistical means to correlate words and contexts is carried out using the text corpus.

The main purpose of this study is weighting the semantic contextual significance of Hindi sentences. That is, investigation of how well the contexts correlate. It is a known cue that the more similar the words are in meaning, the more similar they are in their first order associates.

The proposed approach is to compare Hindi contexts from the largest available Hindi Corpus with the Human judgments of degree of similarity existing between the benchmark word pairs. After having recorded the Human Ratings, the proposed Machine Similarity Measure is defined as shown in **Fig. 3**.

3.3. Data collection with Human Similarity Ratings

To evaluate the proposed similarity measure, firstly the Human Similarity Ratings are recorded. A group of 22 subjects contributed to the Human Similarity Ratings. Human Similarity Ratings is the synonymy judgment given by the subjects on a scale of 0-4 based upon the 'similarity of meaning' in the Hindi word pairs. The 'similarity of meaning' implies the immediate human understanding of the word to make a quick judgment. The subjects were asked to complete a questionnaire to record their judgments. The order of the word pairs was randomized to prevent any biased judgment. The subjects were asked to complete the task in their own leisure time. The subjects were also briefed that they could assign the same values to 1 or more-word pairs. Due to the general property of the word pairs, competent people with technical knowledge of the language were not required to serve as judges. The group of subjects included Graduate and Undergraduate native Hindi Speakers, native Indian Hindi speakers, and Hindi speakers from the Indian Community in Manchester. A diversity of Hindi speakers is chosen for this task, to justify the ratings for the dataset.

The dataset collected with Human Similarity Ratings has the following unique characteristics:

- Randomization: The order of the word pairs is randomized for each subject giving the scores. This is for the prevention of any biased judgment made based on the word order.
- Diversity: The diversity of subjects coming from different age and backgrounds (Graduates, Undergraduates, Businessmen, and Housewives) are chosen for the task.
- Guidelines: The instructions given to the subjects to assign fractional or integer values according to their own accurate perception; and that they can assign the same values to one or more pairs.

A mean score for each word pair ratings is then calculated. The similarity ratings of [0-4] are scaled to [0…1] for comparison of its consistency with the proposed similarity measure. The formula for the proposed similarity measure is explained in the next section.

3.4. Proposed Formula

Different measures have been proposed for calibrating amounts of overlap in the first order associates. A simple formula was adapted as follows:

If A and B are a benchmark word pair, then $\boldsymbol{Ax}$ be a set of words that are extracted from the sentences containing the word $\boldsymbol{A}$ in the Corpus. Let $\boldsymbol{Bx}$ be a set of words that are extracted from the sentences containing the word $\boldsymbol{B}$ in the Corpus.

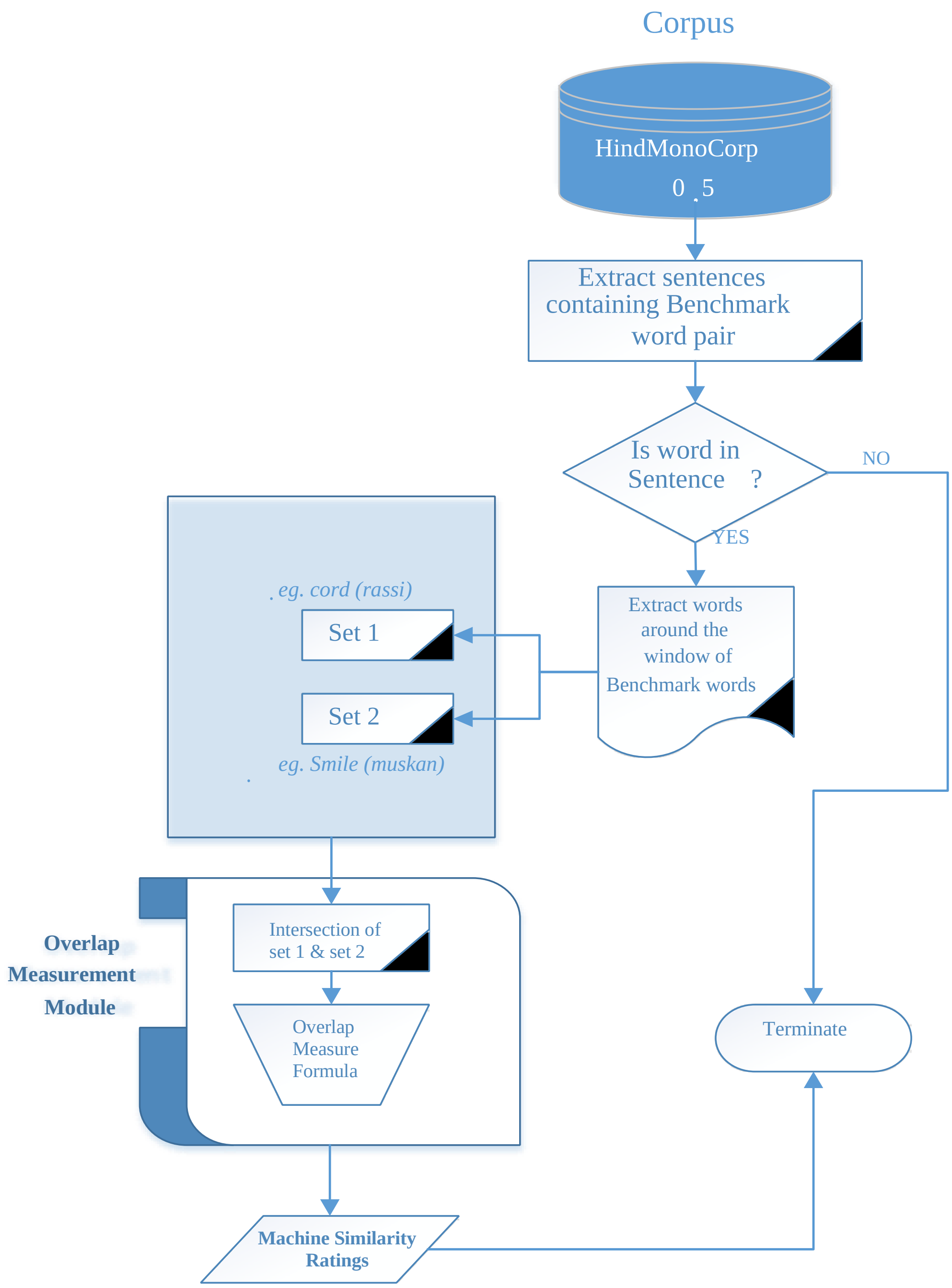

Fig. 3. Proposed Similarity Measure

Then the Overlap in word pair ***A*** & ***B*** is calculated as:

$$\text{Overlap} = \frac{N(Ax \cap B\,x)}{Min\,[N(Ax), N(Bx)]} \quad (3)$$

A = भाई *(brother)*
B = साधु *(monk)*
Sample Calculation for Overlap of the Benchmark word pair:
$Ax = $ [इंसान, बड़ा, जवान, बहन, साहब, विश्वास]
$Bx = $ [ध्यान, बड़ा, महात्मा, इंसान, मंदिरों]
$Ax \cap Bx = $ [बड़ा, इंसान]
$N(Ax) = 6$
$N(Bx) = 5$
$N(Ax \cap Bx) = 2$

$$\text{Overlap} = \frac{N(Ax \cap Bx)}{Min\,[N(Ax), N(Bx)]} = \frac{2}{5} = 0.4$$

The Overlap measure calculated for the above pair is = 0.4

Algorithm of the proposed solution is as follows

Initial data: pair of sixty-five-word pairs $\mathbf{E_x}$ from English benchmark dataset, $\mathbf{H_x} = \emptyset$;
loop
  while $E_x \neq \emptyset$ **do**
    translate $E_i \rightarrow H_i$
    increment $E_i \rightarrow E_{i+1}$
    generate $\mathbf{H_x}$ obtained from translation of $\mathbf{E_x}$
  endwhile;
record human similarity ratings
loop
  while $H_x \neq \emptyset$ **do**
    extract sentence with benchmark word pair A and B
    for word-pair A and B
    if word = = word pair A and B **then**
      if word pair == A **then** add word in set1{} $\mathbf{A_x}$
        else add word in set2{} $\mathbf{B_x}$ **endif;**
      Overlap = $\boldsymbol{N}(A_x \cap B_x) / \boldsymbol{Min}\,[\boldsymbol{N}(A_x), \boldsymbol{N}(B_x)]$
    compare human and machine overlap measure
    else terminate **endif;**
  endwhile;

Let $\mathbf{E_x}$ be a non-empty set of pair of words taken from the English benchmark dataset, where $\mathbf{x}$ range from 1 to 65. $\mathbf{H_x}$ is the empty set that will contain word pairs translated from $\mathbf{E_x}$.

LEMMA 1. At every iteration ***i***, where ***i>0 e***very element of $\mathbf{E_x}$ is translated to H_i generating a set of Hindi words $\mathbf{H_x}$.
Proof. By construction of algorithm, the benchmark dataset $\mathbf{E_x}$ can either be empty or contain word pairs, to begin with, in every iteration, $\mathbf{E_x}$ is checked to be nonempty and perform the translation in Hindi language such that $\mathbf{E_{i+1}}$ is stored in $\mathbf{H_x}$

LEMMA 2. Sets $\mathbf{A_x}$ and $\mathbf{B_x}$ is generated from $\mathbf{H_x}$ i.e. $\mathbf{A_x}$ and $\mathbf{B_x} \in \mathbf{H_x}$ and Overlap measure is calculated from the generated set.

Proof. We first establish that the set $\mathbf{H_x}$ is non empty. At every iteration ***i*,** each sentence containing word pair A and B is extracted from $\mathbf{H_x}$. If the sentence contains the word A or B in the corpus, it is extracted and added to A_x or B_x respectively. In other case, the loop will be terminated. These two non-empty sets are then subjected to the overlap measure formula leading to comparison of calculated human and machine overlap measure

3.5. Implementation Details

Procedure: The machine ratings for the 65-word pairs are calculated using the proposed formula with the help of a Python programming code. Python has a good Unicode support (for the Hindi characters) and large input file handling capabilities. A text file containing around 9.54 million Hindi sentences were used in the code as an input for measuring similarity (file size approx. 2 Gigabytes). The code reads the file and matches the benchmark Hindi word to extract its context around its window. The sets of words are then compared for intersection and the algorithm is applied. **Fig. 3** depicts the detailed working of the code. **Fig. 4** is the code output which gives algorithm measure for a word pair रत्न & गहना (Gem & Jewel).

```
'थ', 'हानहार', 'भारत', 'क्रिकट', 'हुए', 'जस', 'धारण', 'सा', 'भारत', 'क', 'अपन', 'य', 'भारत', 'राजाव', 'प
सहकार', 'अत्यधिक', 'असली', 'रंगीन', 'व्यवसायियों', 'हुए', ' कलर', 'हुआ', 'सम्बंधित', 'जन-साधारण', 'कारण', '
को', 'सूक्ति', 'वाले', '84', 'भारत', 'दिया', 'भारत', 'देना', 'पर', 'होने', 'भारत', 'के', 'भारत', 'के', 'उज्जवल
 'भारत', 'राजीव', 'हैं', 'ऐसे', 'नहीं', 'है।', 'भारत', 'के', 'देने', 'के', 'खेल', 'होता', 'पारदर्शी', 'का', 'खेल',
स्कारों', 'भारत', 'देना', 'माणिक्य', 'को', 'संग्रह', 'ध्वनि', 'भारत', 'दिया', 'का', 'होता', 'भारत', 'से', 'संसारकौन
, 'दोनों', 'भारत', 'शुरु', 'भारत', 'दिए', 'का', 'लहसुनिया', 'के', 'प्राकृतिक', 'धारण', 'यह', 'भारत', 'मंत्री', 'स
भारत', 'से', 'अगीत', 'चौदह', 'का', 'पंचादश', 'भारत', 'क्यों', 'संबंधित', 'पहनकर', 'केवल', 'पुखराज', 'प्रदीप',
', 'कष्टकारी', 'भारत', 'से', 'नवांशहर', 'सिंह', 'एक', 'धन्वंतरि']
1417
Ratan ∩ Gehna = {'अपना', 'खरीदने', 'पहने', 'और', 'वह', 'चन्द्र', 'धर्म', 'ले', 'यही', 'यानी', 'असली', 'बेचने
था', 'तो', 'हल्का', 'यह', 'मैं', 'दान', 'शब्द', 'इत्यादि', 'किसी', 'मिल', 'वह', 'उतार', 'वो', 'आज', 'देने', '
हंगा', 'एवं', 'इतना', 'विशेष', 'नया', 'हुआ', 'कि', 'को', 'के', 'है,', 'पहनकर', 'किस', 'का', 'चंद्र', 'क्यों',
, 'ऐसा', 'है', 'में', 'मिलने', 'सब', 'अपनी', 'एक', 'नही', 'अमूल्य', 'जो', 'लेकर', 'बनवाना', 'कभी', 'हो', '
'कोई', 'बहुमूल्य', 'कछ', 'हर', 'एक-एक', 'जब', 'है।', 'पर', 'रुपी', 'शामिल', 'पहनने', 'वाला', 'नहीं', 'या',
No. of Ratan ∩ Gehna = 113
overlap Measure of Pair 65 = 0.44664031620553357
```

Fig. 4. Machine Overlap Measure

4. Results and Evaluation

Fig. 5 gives both the human and machine similarity ratings calculated. The Human ratings are calculated as the mean judgment given by the participants. Both the relevant judgments are compared for the proposed similarity measure. A variation is seen in the ratings of the English benchmark and Hindi pairs. Both the human and machine ratings give a clear justification of the degree of similarity in the words. The words with low ratings hold the least similarity between them in Hindi. After that, accuracy of the proposed method is calculated.

R&G pair	Hindi Pair	Human Similarity (mean)	Machine Similarity Measure
autograph – shore	स्वाक्षर - किनारे	0	0.00
furnace - implement	भट्टी - लागू	0	0.17
crane – implement	सारस - लागू	0	0.17
cushion – jewel	तकिया - गहना	0.01	0.11
automobile – wizard	वाहन - जादुई	0.01	0.21
rooster – voyage	मुर्गा - जलयात्रा	0.02	0.14
automobile – cushion	वाहन - तकिया	0.02	0.29
asylum – cemetery	शरण - कब्रिस्तान	0.02	0.15
boy – rooster	लड़का - मुर्गा	0.03	0.30
mound – stove	टीला - चूल्हा	0.03	0.10
asylum – fruit	शरण - फल	0.03	0.24
grin – implement	हंसी - लागू	0.04	0.16
noon – string	दोपहर - तार	0.04	0.17
cord - smile	रस्सी - मुस्कान	0.04	0.15
fruit – stove	फल - भट्टी	0.05	0.25
glass – magician	कांच - जादूगर	0.06	0.13
graveyard- madhouse	कब्रिस्तान - पागलख़ाना	0.06	0.18
asylum – monk	शरण - साधु	0.07	0.13
cemetery – mound	कब्रिस्तान - टीला	0.09	0.14
mound – shore	टीला - किनारे	0.11	0.33
boy – sage	लड़का - ऋषि	0.13	0.16
grin – lad	हंसी - बालक	0.13	0.18
brother – monk	भाई - साधु	0.15	0.43
lad – wizard	बालक - जादुई	0.15	0.18
cemetery – woodland	कब्रिस्तान - वनप्रदेश	0.16	0.10
forest – graveyard	वन - कब्रिस्तान	0.16	0.37
monk – slave	साधु - दास	0.17	0.24
implement – tool	लागू - उपकरण	0.18	0.19
coast - forest	तट - वन	0.2	0.26
magician – oracle	जादूगर - आकाशवाणी	0.2	0.10
food – rooster	भोजन - मुर्गा	0.21	0.37
shore – woodland	किनारे - वनप्रदेश	0.22	0.40
sage – wizard	ऋषि - जादुई	0.25	0.13
shore – voyage	किनारे - जलयात्रा	0.26	0.43
coast – hill	तट - पहाड़ी	0.29	0.18
glass - jewel	कांच - गहना	0.29	0.11
monk – oracle	साधु - आकाशवाणी	0.3	0.17
asylum - madhouse	शरण - पागलख़ाना	0.32	0.10
bird – woodland	पक्षी - वनप्रदेश	0.37	0.20
brother – lad	भाई - बालक	0.38	0.43
hill – woodland	पहाड़ी - वनप्रदेश	0.45	0.30
crane - rooster	सारस - मुर्गा	0.47	0.14
food – fruit	भोजन - फल	0.48	0.29
oracle – sage	आकाशवाणी - ऋषि	0.49	0.19
car – journey	कार - यात्रा	0.57	0.25
bird – cock	पक्षी - मुर्गा	0.59	0.26
bird – crane	पक्षी - सारस	0.62	0.35
cord – string	रस्सी - तार	0.66	0.24
journey – voyage	यात्रा - जलयात्रा	0.7	0.84
glass - tumbler	कांच - गिलास	0.74	0.30
gem – jewel	रत्न - गहना	0.74	0.44
magician – wizard	जादूगर - जादुई	0.76	0.13
autograph – signature	स्वाक्षर - हस्ताक्षर	0.81	0.50
serf – slave	कृषक दास - दास	0.83	0.50
hill – mound	पहाड़ी - टीला	0.83	0.30
automobile – car	वाहन - कार	0.85	0.38
furnace – stove	भट्टी - चूल्हा	0.86	0.08
grin – smile	हंसी - मुस्कान	0.87	0.30
forest – woodland	वन - वनप्रदेश	0.87	0.80
coast – shore	तट - किनारे	0.88	0.32
midday - noon	मध्याह्न - दोपहर	0.89	0.55
boy – lad	लड़का - बालक	0.94	0.26
cock – rooster	मुर्गा - मुर्गा	1	1
cushion – pillow	तकिया - तकिया	1	1
cemetery – graveyard	कब्रिस्तान - कब्रिस्तान	1	1

Fig. 5. Similarity Results

To calculate the robustness of the proposed measure, Pearson's correlation coefficient is calculated. In general, a correlation is a method of scrutinizing the relationship between two quanta variables where any correlation coefficient say r is defined as a measure of robustness of similarity between the two variables.

The Pearson's Correlation coefficient achieved for the proposed word similarity method is 0.624. This value is obtained when the three similar word pairs are added with the values 1, as shown in the **Fig. 5**. When these three-word pairs are removed, the correlation coefficient obtained is 0.514. The human and machine similarity ratings are plotted on the graph.

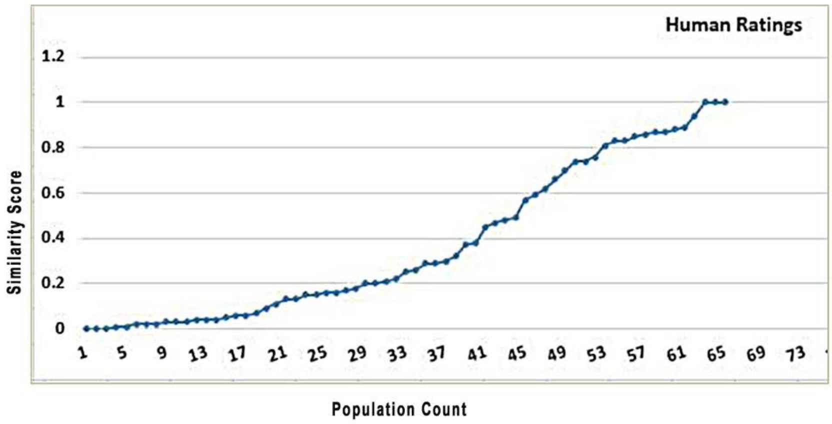

Fig. 6. Human Ratings

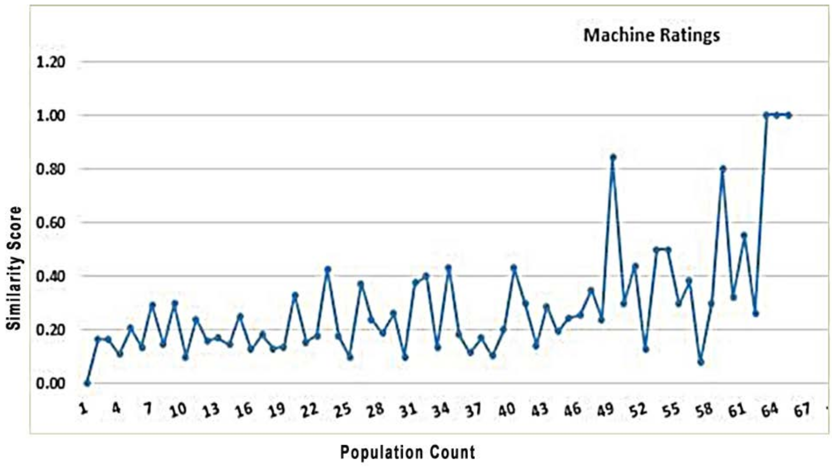

Fig. 6. Machine Ratings

5. Conclusion and Future Work

A word semantic similarity measure for Hindi is proposed in this work. Firstly, a detailed review of the available resources and knowledge bases for Hindi is performed, and then an overlap measure is adapted to calculate the relation between a word and its contexts. A corpus is a good lexical resource that contains the actual language usage by humans. A corpus can thus be applied to a diverse range of applications in Natural Language Processing. For the proposed similarity measure, the English benchmark dataset is translated to Hindi. A complete description of ethical translation has been provided. A similarity measure has been proposed which calculates the amount of contextual overlap between the word pairs. The human similarity ratings are the mean of the judgments given by 22 subjects or the speakers of the Hindi language. The proposed similarity measure is then compared to human similarity ratings. A good coefficient correlation is achieved for the proposed measure.

The main advantage of the work proposed in this paper is the faithful translation of the benchmark dataset which can be adapted for various applications of Natural Language Processing. The proposed measure can also be adapted for various other languages.

The Hindi language has a long way to go in terms of dialogue and conversational agents. But due to the simple structure and understanding, various methods can be adapted and applied to attain better linguistic learning of the language. Future work includes consideration of the post prepositions and prepositions in the Hindi Language. Consideration of the cumulative frequency of words in the language can be a key to better measures in the future. More varied datasets that include sentence similarity can also be adapted for the Hindi language.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] I. Y. Junghare, "'Avatar': Heavenly Descent," *International Journal of Diversity in Organisations, Communities & Nations*, vol. 11, no. 4, pp. 121-137, 2012. Article (CrossRef Link)

[2] M. Choice and P.M.D.M. Power, "India Human Development Survey," 2016. Article (CrossRef Link)

[3] Taieb, Mohamed Ali Hadj, Torsten Zesch, and Mohamed Ben Aouicha, "A survey of semantic relatedness evaluation datasets and procedures," *Artificial Intelligence Review*, vol. 53, no. 6, 4407-4448, 2020. Article (CrossRef Link)

[4] Sebastiani, Fabrizio, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1-47, 2002. Article (CrossRef Link)

[5] Z. Akhtar, J. W. Lee, M. A. Khan, M. Sharif, S. A. Khan, and N. Riaz, "Optical character recognition (OCR) using partial least square (PLS) based feature reduction: an application to artificial intelligence for biometric identification," *Journal of Enterprise Information Management*, Jul. 31 2020. Article (CrossRef Link)

[6] M. Sharif, M. A. Khan, M. Faisal, M. Yasmin, and S. L. Fernandes, "A framework for offline signature verification system: Best features selection approach," *Pattern Recognition Letters*, vol. 139, pp. 50-59, 2020. Article (CrossRef Link)

[7] F. E. Batool, M. Attique, M. Sharif, K. Javed, M. Nazir, A. A. Abbasi, et al., "Offline signature verification system: a novel technique of fusion of GLCM and geometric features using SVM," *Multimedia Tools and Applications*, pp. 1-20, Apr. 2020. Article (CrossRef Link)

[8] S. Kopp, L. Gesellensetter, N. C. Krämer, and I. Wachsmuth, "A conversational agent as museum guide–design and evaluation of a real-world application," in *Proc. of International workshop on intelligent virtual agents*, pp. 329-343, 2005. Article (CrossRef Link)

[9] T. Saba, M. Bashardoost, H. Kolivand, M. S. M. Rahim, A. Rehman, and M. A. Khan, "Enhancing fragility of zero-based text watermarking utilizing effective characters list," *Multimedia Tools and Applications*, vol. 79, pp. 341-354, 2020. Article (CrossRef Link)

[10] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," *Communications of the ACM*, vol. 8, pp. 627-633, 1965. Article (CrossRef Link)

[11] J. O'shea, Z. Bandar, and K. Crockett, "A new benchmark dataset with production methodology for short text semantic similarity algorithms," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 10, pp. 1-63, 2014. Article (CrossRef Link)

[12] N. Adel, K. Crockett, A. Crispin, J. P. Carvalho, and D. Chandran, "Human Hedge Perception–and its Application in Fuzzy Semantic Similarity Measures," in *Proc. of IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1-7, 2019. Article (CrossRef Link)

[13] R. Javadzadeh, M. Zahedi, and M. RAHIMI, "Sentence similarity using weighted path and similarity matrices," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 27, pp. 3779-3790, 2019. Article (CrossRef Link)

[14] J. Guo, H. He, T. He, L. Lausen, M. Li, H. Lin, et al., "Gluoncv and gluonnlp: Deep learning in computer vision and natural language processing," *Journal of Machine Learning Research*, vol. 21, pp. 1-7, 2020. Article (CrossRef Link)

[15] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 604-624, 2021. Article (CrossRef Link)

[16] V. Sorin, Y. Barash, E. Konen, and E. Klang, "Deep Learning for Natural Language Processing in Radiology—Fundamentals and a Systematic Review," *Journal of the American College of Radiology*, vol. 17, no. 5, pp. 639-648, 2020. Article (CrossRef Link)

[17] Y. Li, Z. A. Bandar, and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources," *IEEE Transactions on knowledge and data engineering*, vol. 15, pp. 871-882, 2003. Article (CrossRef Link).

[18] I. Atoum, "A novel framework for measuring software quality-in-use based on semantic similarity and sentiment analysis of software reviews," *Journal of King Saud University-Computer and Information Sciences*, vol. 32, pp. 113-125, 2020. Article (CrossRef Link)

[19] S. H. Wasti, M. J. Hussain, G. Huang, A. Akram, Y. Jiang, and Y. Tang, "Assessing semantic similarity between concepts: A weighted-feature-based approach," *Concurrency and Computation: Practice and Experience*, vol. 32, no. 7, p. e5594, 2020. Article (CrossRef Link)

[20] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," *Aaai*, vol. 6, no. 2006, pp. 775-780, 2006. Article (CrossRef Link)

[21] F. Smarandache, M. Colhon, Ş. Vlăduţescu, and X. Negrea, "Word-level neutrosophic sentiment similarity," *Applied Soft Computing*, vol. 80, pp. 167-176, 2019. Article (CrossRef Link)

[22] H. K. Azad and A. Deepak, "A novel model for query expansion using pseudo-relevant web knowledge," *arXiv preprint arXiv:1908.10193*, 2019. Article (CrossRef Link)

[23] S. Badodekar, "Translation resources, services and tools for Indian languages," *Computer Science and Engineering Department, Indian Institute of Technology, Mumbai*, vol. 400019, 2003. Article (CrossRef Link)

[24] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32-73, 2017. Article (CrossRef Link)

[25] A. Jain, S. Vij, and O. Castillo, "Hindi Query Expansion based on Semantic Importance of Hindi WordNet Relations and Fuzzy Graph Connectivity Measures," *Computación y Sistemas*, vol. 23, no. 4, pp. 1337–1355, 2019. Article (CrossRef Link)

[26] G. Jain and D. Lobiyal, "Word Sense Disambiguation of Hindi Text using Fuzzified Semantic Relations and Fuzzy Hindi WordNet," in *Proc. of 9th Int. Conf. on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 494-497, 2019. Article (CrossRef Link)

[27] Pandit, Rajat, Saptarshi Sengupta, Sudip Kumar Naskar, Niladri Sekhar Dash, and Mohini Mohan Sardar, "Improving Semantic Similarity with Cross-Lingual Resources: A Study in Bangla—A Low Resourced Language," *Informatics*, vol. 6, no. 2, p. 19, 2019. Article (CrossRef Link)

[28] M. Sinha, M. Reddy, and P. Bhattacharyya, "An approach towards construction and application of multilingual indo-wordnet," in *Proc. of 3rd Global Wordnet Conference (GWC 06)*, Jeju Island, Korea, 2006. Article (CrossRef Link)

[29] N. Mishra and A. Mishra, "Part of speech tagging for Hindi corpus," in *Proc. of Intl. Conf. Communication Systems and Network Technologies (CSNT)*, pp. 554-558, 2011. Article (CrossRef Link)

Farah Younas has done her MSc Advanced Computer Science from Manchester Metropolitan University, Manchester, England. She is currently a Lecturer of Computer Science in the department of Computing at Shaheed Zulfikar Ali Bhutto Institute of Science and Technology, Islamabad, Pakistan. She has over 3 years' experience in Spring MVC, Hibernate, JAVA, Python, PHP, wordpress and Web Application Development. Her research interest includes Natural Language Processing, Data Mining, Machine Learning, Information extraction from big data.

Jumana Nadir is a Masters of Software Engineering student at the San Jose State University, San Jose, CA, USA. She is currently working as a part time Machine Learning Engineer at DataPrime Inc., California, USA. She has over 2 years' experience in Data Science, building Machine Learning pipelines, Python, and Java. Her research interest includes Natural Language Processing, Data Mining and Deep Learning and Artificial Intelligence.

Muhammad Usman has completed PhD in Computer & Information Sciences from Auckland University of Technology, New Zealand. He is currently an Associate Professor of Computer Science in the department of Computing at Shaheed Zulfikar Ali Bhutto Institute of Science and Technology, Islamabad, Pakistan. His research interests include Data Mining, Data Warehousing, OLAP, Business Intelligence and Knowledge discovery. He is currently researching in the novel methods and techniques for the seamless integration of Data Mining and Data Warehousing technologies. He has published in international journals and conference proceedings, and he has served as reviewer for several premier journals and conferences.

Sajid Ali Khan has completed PhD in Software Engineeing from Islamic University Islamabad, Pakistan. He is currently an Assist Professor of Software Engineering in the department of SE at Foundation University Islamabad Pakistan. His research interests include image processing, machine learning and pattern recognition.He is reviewer of many well-repute journals such as MTAP, IEEE Access, and name a few more.

Muhammad Attique has completed PhD in Computer Science from COMSATS University Islamabad, Wah Campus Pakistan. He is currently an Lecturer of Computer Science in the department of Computer Sciuence at HITEC University Taxila Pakistan. His research interests Deep Learning, Computer Vision, and medicla imaging. He published more than 140 articles of h-index 30 and total citations above 2350. He is reviewers of many well reputed journals such as IEEE Journal of Biomedical and Health Informatics, Neurocomputiung, Information Sciences, and name a few more.

Seifedine Kadry is currently an Professor at Department of Applied Data Science, Noroff University College, Norway. Previously, he is an associate professor at Beirut Arab University, Lebanon, and was previously a Professor with American University of the Middle East in Kuwait. He serves as Editor-in-Chief of the Research Journal of Mathematics and Statistics and the ARPN Journal of Systems and Software. He worked as Head of Software Support and Analysis Unit of First National Bank where he designed and implemented the data warehouse and business intelligence. In addition, he has published several books and is the author of more than 100 papers on applied math, computer science, and stochastic systems in peer-reviewed journals. At present his research focuses on smart learning in smart cities, social network analysis, and E-systems. He received a PhD in computing 2007 from the Blaise Pascal University (Clermont-II) -Clermont-Ferrand in France. He is an IEEE senior member.

Yunyoung Nam received the B.S., M.S., and Ph.D. degrees in computer engineering from Ajou University, Korea in 2001, 2003, and 2007 respectively. He was a Senior Researcher with the Center of Excellence in Ubiquitous System, Stony Brook University, Stony Brook, NY, USA, from 2007 to 2010, where he was a Postdoctoral Researcher, from 2009 to 2013. He was a Research Professor with Ajou University, from 2010 to 2011. He was a Postdoctoral Fellow with the Worcester Polytechnic Institute, Worcester, MA, USA, from 2013 to 2014. He was the Director of the ICT Convergence Rehabilitation Engineering Research Center, Soonchunhyang University, from 2017 to 2020. He has been the Director of the ICT Convergence Research Center, Soonchunhyang University, since 2020, where he is currently an Assistant Professor with the Department of Computer Science and Engineering. His research interests include multimedia database, ubiquitous computing, image processing, pattern recognition, context-awareness, conflict resolution, wearable computing, intelligent video surveillance, cloud computing, biomedical signal processing, rehabilitation, and healthcare systems.