Text Summarization Technique for Punjabi Language Using Neural Networks

Arti Jain[1], Anuja Arora[1], Divakar Yadav[2], Jorge Morato[3], and Amanpreet Kaur[1]
[1]Department of Computer Science and Engineering, Jaypee Institute of Information Technology, India
[2]Department of Computer Science and Engineering, National Institute of Information Technology, India
[3]Department of Computer Science and Engineering, Universidad Carlos III de Madrid, Spain

Abstract: *In the contemporary world, utilization of digital content has risen exponentially. For example, newspaper and web articles, status updates, advertisements etc. have become an integral part of our daily routine. Thus, there is a need to build an automated system to summarize such large documents of text in order to save time and effort. Although, there are summarizers for languages such as English since the work has started in the 1950s and at present has led it up to a matured stage but there are several languages that still need special attention such as Punjabi language. The Punjabi language is highly rich in morphological structure as compared to English and other foreign languages. In this work, we provide three phase extractive summarization methodology using neural networks. It induces compendious summary of Punjabi single text document. The methodology incorporates pre-processing phase that cleans the text; processing phase that extracts statistical and linguistic features; and classification phase. The classification based neural network applies an activation function-sigmoid and weighted error reduction-gradient descent optimization to generate the resultant output summary. The proposed summarization system is applied over monolingual Punjabi text corpus from Indian languages corpora initiative phase-II. The precision, recall and F-measure are achieved as 90.0%, 89.28% an 89.65% respectively which is reasonably good in comparison to the performance of other existing Indian languages' summarizers.*

Keywords: *Extractive method, Indian languages corpora initiative, natural language processing, neural networks, Punjabi language, text summarization.*

Received May 31, 2020; accept January 6, 2021
https://doi.org/10.34028/iajit/18/6/8

1. Introduction

In contemporary days, exploitation of digital information has risen considerably such as newspaper and web articles, status updates, tweets [21] and advertisements that have become a part of our daily basis routine. Due to the digitized information overload over websites and web portals, there is a dire need to build an automated summarization system which yields textual summary in a meaningful and compendious way.

Natural Language Processing (NLP) [19] is deemed to enable the computer to understand, analyze and interpret human languages. Text Summarization (TS) [26, 49] is a field of NLP which is not a neophyte subject but is under evolution for more than four decades. There are two paradigms in the text summarization [10, 46]-extractive and abstractive summarization. Extractive summarization [8, 11] selects important sentences as text snippets from the original text, weigh them with statistical features and linguistic measures. In short, it is a binary classification of sentence, depending upon whether sentence is included in the summary or not. Abstractive summarization [43] tries to understand the original text where output includes paraphrasing, generalization and real-world knowledge to rephrase the text in fewer words. TS based research is easily available for the English language e.g., Text REtrieval Conference (TREC) tracks [3]-temporal summarization track, MultiLing workshop at text analysis conference are to name a few. In 2016, both tracks-temporal summarization track and microblog track are merged in real-time summarization [35]. The groundbreaking studies-See *et al*. [43], Liu and Lapata [36], and Aries *et al*. [2] are worth mentioning. These studies show that despite great advances in the text summarization task, there is a need of pursuing research in this area due to the current information growth. Apart from this, there are morphological rich languages such as Punjabi where text summarization process is still in premature stage. There are 125 million Punjabi speakers, not only in India and Pakistan, but in many other countries all over the world. Literacy rate in the Punjab has grown 6 points in 10 years, now is 75%. However, the Punjabi language has specific issues which hinder the summarization process, like: postpositions, lack of standardization, no capitalization, complex morphology, fast evolution, different dialects, and paucity of linguistic resources.

1.1. Postpositions

The Punjabi language has postpositions rather than

prepositions and paraphrases, e.g., ਨਸ਼ੇ ਦੀ ਲੱਤ ਲੱਗਣਾ *Naśē dī lata lagaṇā* "addiction to drugs" vs. ਨਸ਼ੇੜੀ *Naśēṛī* "drug addict".

1.2. Lack of Standardization

The Punjabi language is codified in different scripts, mainly-Gurmukhi and Shahmukhi. Even within the same script there are different spellings due to the usage of diacritics, as in Table 1.

Table 1. Sample punjabi diacritics with examples.

Diacritic	Top/Foot Character	Example
addhak (ੱ)	Top	ਪੱਤਾ *pattā* "leaf"
tippī (ੰ)	Top	ਮੁੰਹ *mũ:* "mouth"
bindī (ਂ)	Top	ਬਾਂਹ *bắh* "arm"
(◌ਵ)	Foot	ਸ੍ਵਰਗ *svāragā* "heaven"
(◌ਹ)	Foot	ਮੀਂਹ *mĩ́* "rain"

1.3. No Capitalization

The Punjabi language has no concept of capitalization within the proper nouns.

1.4. Complex Morphology

The Punjabi language has complex morphological structure (root complexity and syntactic diversity).

1.5. Fast Evolution

The Punjabi language incorporates several English nouns into it (e.g., technology ਤਕਨਾਲੋਜੀ takanālōjī).

1.6. Different Dialects

The Punjabi language has many local variations and dialects [28, 44].

1.7. Paucity of Linguistic Resources

The Punjabi linguistic resources are built from limited resources, as in Table 2. The Punjabi NLP tools dates back from eight-to-ten years ago, and are developed from fewer resources. For example, Gupta and Lehal [14] have developed the Punjabi resources using newspaper-Ajit.

Table 2. Punjabi resources with references.

Punjabi Resources	Reference(s)
Stop-words lists	Kaur and Saini [29]; Gupta and Lehal [14]
Ontology and WordNet	Kaur and Sharma [30]; Kaur *et al*. [31]; Krail and Gupta [32]
Stemming tools	Gupta and Lehal [11]
Normalization	Gupta [12]
Part of speech tagging	Gill *et al*. [6]; Gupta and Lehal [14]
Named entity recognition	Kaur *et al*. [27]; Gupta and Lehal [15]
Gazetteers	Gupta and Lehal [13]

In the survey conducted by Aries *et al*. [2], problem with the lack of resources in some languages is mentioned. It is common to apply summarization methods on languages such as English. Here, in the present work, an extractive summarization using three-phase methodology is proposed on another problem domain i.e., Summarization task for the Punjabi language. The proposed methodology involves preprocessing, processing and classification phases which induces meaningful short summary over Unicode encoded monolingual Punjabi text corpus. The preprocessing phase cleans the Punjabi text; processing phase extracts the statistical and linguistic features; and classification based Neural Network (NN) undergoes weight inclusion during the forward pass and weight updation during the backward pass until convergence or suitable number of iterations is accomplished. It is worth mentioning that in comparison to other techniques [9, 10], the neural network does not impose restriction on the input variables. Previously, the NN is useful in speech recognition [48], cancer detection [38], stock prices [18], and language modeling [50] etc., In other words, NN is well-suited for data with high volatility and non-constant variance, able to learn hidden relationships that too without imposing fixed relationships within data.

The highest scored sentences are added to the generated summary while achieving precision-90.02%, recall-89.28%, and F-measure-89.65% respectively which is quite competitive w.r.to existing summarizers for other Indian languages' such as Bengali, Hindi, Gujarati, Urdu, Kannada. To the best of our knowledge, no work using the proposed methodology has ever been considered so far for the Punjabi. This way it is a novel work.

Rest of the paper is outlined as follows. Section 2 discusses the related work. Section 3 mentions the proposed methodology. Section 4 illustrates experimental setup, Punjabi dataset and results. Section 5 concludes the paper.

2. Related Work

Gupta and Lehal [10] have surveyed extractive text summarization techniques while discussing features such as keyword, title word, sentence location, sentence length, proper noun, upper-case word, cue-phrase, sentence-to-sentence cohesion etc., The general extractive summarization methods include- cluster based, graph theoretic, machine learning, latent semantic analysis, neural networks, fuzzy logic, regression and query based. Gupta and Lehal [14] have detailed a pre-processing phase within the Punjabi summarization task. The pre-processing sub-phases involve- elimination of Punjabi stop-words, Punjabi stemmer for nouns, normalization of Punjabi nouns, and elimination of duplicate Punjabi sentences. The pre-processing is done on 50 Punjabi news documents and stories, comprising

of 11.29 million words from the Punjabi news daily-Ajit with an efficiency gain of 32% at 50% compression rate. Gupta and Lehal [13] have worked on extractive summarizer for single document based Punjabi text. The statistical features are- keywords, sentence length, and numbered data. The linguistic features are- Punjabi headlines and next lines, Punjabi nouns and proper nouns, Punjabi cue phrases and Punjabi title keywords. Based on the variety of features, fuzzy scores to the Punjabi sentences are executed which is followed by the regression to calculate the feature weights. The high scored sentences are selected in a particular order, within the generated summary. Gupta and Kaur [9] have implemented support vector machine for Punjabi summarization using conceptual, statistical and linguistic features.

Apart from Punjabi, other languages such as English and Hindi too perform text summarization. Gupta [8] has worked with hybrid algorithm over 30 Hindi-Punjabi documents for TS task. The author has combined nine features as are suggested by Centre for Development of Advanced Computing (C-DAC), Noida, India. These features are- key phrase extraction, font, noun-verb extraction, position, cue-phrase, negative keyword, named entity, relative length, and numbered data. The mathematical regression is applied over features score and sentences are scored from the feature weight equations. It has achieved F-measure of 92.56%. Kumar *et al*. [34] have used a graph-based approach for the Hindi summarization where sentences are ranked based on the words frequency and semantic analysis. Kumar and Yadav [33] have worked with the thematic approach to select significant sentences for the Hindi TS. The stop-words elimination and stemming process are executed before selection of the thematic words. The system is tested using expert game and has achieved an accuracy of 85%. Singh *et al*. [45] have presented a bilingual, unsupervised, automatic text summarization using deep learning. They have extracted 11 features to generate a feature matrix. To improve accuracy, the matrix is passed through the restricted boltzmann machine and a reduced version of the document is generated without losing the important information and has achieved accuracy of about 85%. Dalal and Malik [5] have summarized the Hindi document using particle swarm optimization. The subject-object-verb triplets are extracted to construct a semantic graph of the document and to obtain the desired summary. Gulati and Sawarkar [7] have built a fuzzy inference engine to summarize online Hindi news articles on sports and politics. They have used 11 features and have achieved 73% precision. Dalal and Malik [4] have worked with bio-inspired computing for the Hindi summarization over Cross Language Indian News Story Search (CLINSS) corpus. The corpus consists of Hindi news articles related to politics, events, sports, history and stories etc. They have achieved precision (42.86%), recall (60%), F-measure (50.01%) and G-score (50.71%) respectively. See *et al*. [43] have used hybrid pointer-generator architecture to copy words from source text via pointing, and coverage to track what is summarized to discourage repetition. The model is applied to long text dataset from CNN/Daily Mail, outperforming by at least 2 Recall-Oriented Understudy for Gisting Evaluation (ROUGE) points. Liu and Lapata [36] have showed that a bidirectional encoder representation from transformer is applied to TS. The extractive model is constructed on top of the encoder while stacking several inter-sentences transformer layers. The experiments are conducted over three datasets-Cable News Network (CNN)/Daily Mail, New York Times (NYT) and XSum. Mohd *et al*. [37] have preserved text semantics as feature for the summarization task using ROUGE over DUC'07. Prudhvi *et al*. [40] have applied unsupervised learning, cosine similarity [17], and rank algorithm for the text summarization.

Based on the literature review, the following Research Gaps (RGs) are identified for the Punjabi text summarization task which motivates us to work in this direction.

- *RG* 1: There is a lack of resources that are useful for the pre-processing phase.

The pre-processing involves text cleaning tasks-removing stop-words, stemming, normalization and elimination of duplicate sentences. This task for the Punjabi text has been performed by Gupta and Lehal [14] for genres, like news. There is a dire need to assimilate them from multiple resources.

- *RG* 2: There is a lack of certain statistical and linguistic features that are beneficial for the processing phase.

Previous studies have worked upon various conceptual, statistical and linguistic features [9, 10, 13]. But other vital features such as Term Frequency-Inverse Sentence Frequency (TF-ISF) and Named Entity Recognition (NER) [20, 22] are to be investigated for the summarization task.

- *RG* 3: Exploration of a classification method is required whose implementation provides effective results during the classification phase.

Previous studies have applied [9, 10, 13, 14]-cluster based, graph theoretic, fuzzy logic, regression model, query based, genetic algorithm, feed-forward neural networks and Gaussian mixture model for the Punjabi summarization task. The classification based neural networks [51] is to be explored for weight inclusion and weight updation for features until either convergence or suitable number of iterations is accomplished.

- *RG* 4: There is a great need for a standard dataset for the Punjabi text summarization task. This dataset has to be richer and representative of language for

the experimentation purpose.

Previous studies have mainly experimented with the Punjabi news Daily-Ajit [14]. There is a keen necessity of standard Punjabi dataset for the summarizations task.

In order to overcome the above stated research gaps, we have proposed an extractive Punjabi text summarization methodology with the following research objectives:

- *RO* 1: To embed multiple resources- stemmer, normalizer and elimination of stop-words for the Punjabi text at one go.
- *RO* 2: To include TF-ISF and NER features during processing phase of the Punjabi summarization task.
- *RO* 3: To select classification based neural networks for summary generation of the Punjabi text.
- *RO* 4: To consider standard monolingual Punjabi dataset for the experimentation purpose.

In order to fulfill the above stated research objectives, the proposed Punjabi text summarization methodology has the following main Research Contributions (RCs):

- *RC* 1: The pre-processing phase [16] involves cleaning of the Punjabi text via removal of punctuation, input restriction [23], sentence tokenization, word tokenization, stemming [14], normalization [12] and stop-words elimination [28].
- *RC* 2: The cleaned Punjabi text undergoes the processing phase which extracts statistical and linguistic features and calculates scores of the sentences. The distinguished Punjabi features are- TF-ISF, headlines and next lines, NER, cue-phrases, nouns and Common Punjabi-English Nouns (CPEN).
- *RC* 3: The classification based neural networks is applied for summary evaluation which induces those Punjabi sentences that are relevant to the summary, also computes the precision, recall and F-measure of the proposed system. The NN is able to learn non-linear, complex relationships among sentences that persist within a language. The neural network learns from initial sentences and their relationships, then becomes capable to generalize, and so predicts over unseen sentences.
- *RC* 4: The Punjabi dataset is collected as a monolingual Punjabi text corpus under the Indian Languages Corpora Initiative Phase-II (ILCI Phase-II). The ILCI project is initiated by the Ministry of Electronics and Information Technology (MeitY), Government of India.

3. Proposed Methodology

The proposed methodology constitutes- pre-processing, processing and classification phases respectively.

3.1. Pre-Processing Phase

In the pre-processing, an initial illustration over the textual data is marked through the given tasks.

3.1.1. Removal of Punctuation

Punctuations such as- . “” : are eliminated from the Punjabi sentences.

3.1.2. Input Restriction

Majority of the text has to be written in the Punjabi. So, the length of the Punjabi characters should not be less than 80% of the total.

3.1.3. Sentence Tokenization

Presence of sentence indicators e.g., । ? ! are responsible for sentence boundary in the Punjabi text. For example, the vertical bar (।) indicates end of a punjabi sentence.

3.1.4. Word Tokenization

Each tokenized Punjabi sentence can be tokenized into words for easing the tasks such as elimination of stop-words, extraction of features etc.

3.1.5. Stemming

The stemming task marks the Punjabi words into their basic form. The Punjabi stemmer that is built by Gupta and Lehal [14] is taken into consideration which has an accuracy of 87.37%. An example indicating different inflectional forms of a Punjabi word ਸੋਹਣਾ *sōhṇā* “beautiful” is given in Table 3.

Table 3. Inflectional forms of a punjabi word.

Word	Masculine/ Feminine	Inflectional Form	Singular/ Plural
ਸੋਹਣਾ *sōhṇā* “beautiful”	Masculine	ਸੋਹਣਾ *sōhṇā* “beautiful”	Singular
		ਸੋਹਣੇ *sōhṇē* “beautiful”	Plural
ਸੋਹਣੀ *sōhṇī* “beautiful”	Feminine	ਸੋਹਣੀ *sōhṇī* “beautiful”	Singular
		ਸੋਹਣੀਆਂ *sōhṇīāṃ* “beautiful”	Plural

3.1.6. Normalization

There are many spelling variations in the Punjabi. To overcome the same, Punjabi noun morph is normalized using Punjabi normalizer that is built by Gupta [12]. For example: ਹਨੁਮਾਨਗੜ੍ਹ *Hanūmānagaṛha* “Hanumangarh” is also written as ਹਨੁਮਾਨਗੜ੍ *Hanūmānagaṛ* “Hanumangarh”. And, ਖ਼ਿਆਲ *khaiāl* “idea” is also written as ਖਿਆਲ *khaiāl* “idea”. So, the words are to be normalized.

3.1.7. Elimination of Stop-Words

Stop-words such as ਦੇ *dē* “of”, ਵਿਚ *vica* “in the”, ਨਾਲ

nāla "with", ਹੈ *hai* "is" and so on, do not convey significant meaning, so are eliminated from the Punjabi text. There are 184 stemmed Gurmukhi stop-words, as is suggested by Kaur and Saini [29].

3.2. Processing Phase

In the processing phase, different statistical and linguistic features [23] are extracted as follows:

3.2.1. Term Frequency-Inverse Sentence Frequency

TF-ISF [1] is the most commonly used feature in NLP to extract important keywords from a text, as in Equation (1).

$TF(t)$: word frequency within Punjabi sentence
$ISF(t)$: $log\ (N/N_t)$
N : sentences count with Punjabi text
N_t : sentences count having the word t

$$TS\text{-}ISF(t) = TF(t) * ISF(t) \quad (1)$$

3.2.2. Headlines and Next Lines

A headline of a text document is an important feature which conveys core theme of the Punjabi text e.g.,
ਆਈ. ਸੀ. ਐੱਸ. ਈ 10ਵੀਂ ਤੇ 12ਵੀਂ ਦੇ ਨਤੀਜਿਆਂ ਦਾ ਐਲਾਨ ਅੱਜ
(*I.C.S.E results of X and XII are announced today*)

ਚੰਡੀਗੜ੍ਹ (ਰਸ਼ਮੀ): *Caḍīgaṛha (raśamī)* "Chandigarh (Rashmi)" Chandigarh (Rashmi) is the line next to the headline which interprets the location and name of the author of the above stated Punjabi sentence.

3.2.3. Named Entity Recognition

NER [24] extracts named entities such as names of person, locations, organizations etc., from the text. Extraction of punjabi named entities include rule-based methodology and gazetteers. Different Punjabi gazetteer lists [13]-prefix, suffix, middle and last names, list of names etc., are used to check whether a given word is named entity or not. For example:
ਡੋਨਲਡ ਟਰੰਪ *Ḍōnalaḍa ṭarapa* "Donald Trump" (Person name) ਐਪਲ ਇਨਕਾਰਪੋਰੇਟਿਡ *Aipala inakārapōrēṭiḍa* "Apple Incorporated" (Organization name).

3.2.4. Cue-Phrases

Presence of cue-phrases [14] in the sentences is emphasized as they have important meaning to tell. The sentences which contain cue-phrases are considered more weight-age instead of without them. For example: ਅੰਤਵਿੱਚ *Atavica* "in the end", ਸੰਖੇਪਵਿੱਚ *Sakhēpavica* "in brief".

3.2.5. Noun and Common Punjabi-English Nouns

Nouns have higher weight-age and nowadays it is common that some English nouns are written in the Punjabi too. The accuracy of Punjabi nouns and CPEN identification are 98.43% and 95.12% respectively, as is stated by Gupta and Lehal [13]. For example: Technology is written in Punjabi as: ਟੇਕਨਾਲੋਜੀ (*Ṭēkanālōjī*).

3.3. Classification Phase

In this work, classification based neural network learns the Punjabi sentences those are inclusive within the summary. To do so, NN based backpropagation is used to discover the patterns which comprises of 5 input neurons, 1 hidden layer and 1 output layer respectively. In the input layer there are five features-TF-ISF, headlines and next lines, NER, cue-phrases and CPEN that are extracted from the processing phase and weights are assigned to each neuron. The hidden layer with bias computes the sum of the weighted features, and then weighted connection with sigmoid activation function flows into the output neuron. The output layer with a bias calculates the output, and error is propagated back to the hidden layer. The gradient descent optimization propagates error while updating weights until the generated output approximates the targeted output summary (Figure 1).

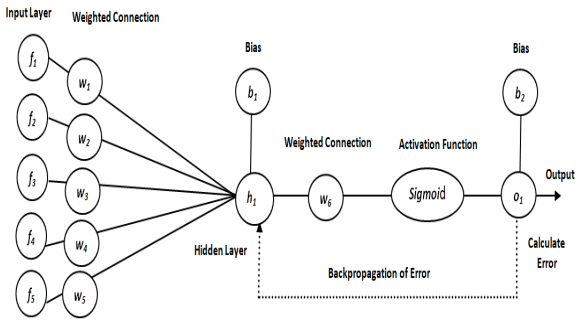

Figure 1. Neural networks for Punjabi text summarization.

The backpropagation neural networks work in forward and backward passes. Each one of them is detailed here.

3.3.1. Forward Pass

In the forward pass, net input at the hidden layer (net_{h1}) is calculated as the sum of feature weights that are obtained from the processing phase and bias (b_1), as in Equation (2).

$$net_{h1} = w_1f_1 + w_2f_2 + w_3f_3 + w_4f_4 + w_5f_5 + b_1 \quad (2)$$

Here, f_1, f_2, f_3, f_4 *and* f_5 are features, and w_1, w_2, w_3, w_4 and w_5 are the weights that are assigned to the feature set. The output of the hidden layer (out_{h1}) is computed using sigmoid function, as in Equation (3).

$$out_{h1} = \frac{1}{1+e^{-net_{h1}}} \quad (3)$$

Then the net input at output layer (net_{o1}) is calculated as the sum of weighted connection (w_6) to output of the hidden layer and bias (b_2), as in Equation (4).

$$net_{o1} = w_6 * out_{h1} + b_2 * 1 \quad (4)$$

Thus, the computed output at output layer (out_{o1}) using sigmoid function is observed, as in Equation (5).

$$out_{o1} = \frac{1}{1+e^{-net_{o1}}} \tag{5}$$

However, the computed output is compared with the target output to calculate the value of error (E), as in Equation (6).

$$E = \frac{1}{2}(t\arg et_{o1} - out_{o1})^2 \tag{6}$$

3.3.2. Backward Pass

In the backward pass, the error is fed back through the network to adjust weights of each connection and reduces the error by a small amount. At the output layer, how much change in w_6 affects the error is to be known. For this, derivative of error w.r.to weighted connection (w_6) is computed, as in Equation (7).

$$\frac{\partial E}{\partial w_6} = \frac{\partial E}{\partial out_{o1}} * \frac{\partial out_{o1}}{\partial net_{o1}} * \frac{\partial net_{o1}}{\partial w_6} \tag{7}$$

However, using Equation (6) $\frac{\partial E}{\partial out_{o1}}$ is obtained, asEquation in (8)

$$\frac{\partial E}{\partial out_{o1}} = -(t\arg et_{o1} - out_{o1}) \tag{8}$$

Also, using Equation (5) $\frac{\partial out_{o1}}{\partial net_{o1}}$ is obtained, as in Equation (9)

$$\frac{\partial out_{o1}}{\partial net_{o1}} = out_{o1} * (1 - out_{o1}) \tag{9}$$

And, using Equation (4) $\frac{\partial net_{o1}}{\partial w_6}$ is obtained, as in Equation (10)

$$\frac{\partial net_{o1}}{\partial w_6} = out_{h1} \tag{10}$$

Thus, using Equations (7-10) $\frac{\partial E}{\partial w_6}$ becomes, as in Equation (11)

$$\frac{\partial E}{\partial w_6} = -(t\arg et_{o1} - out_{o1}) * out_{o1}(1 - out_{o1}) * out_{h1} \tag{11}$$

Now, use the delta δ_{o1}, as in Equation (12)

$$\begin{aligned} \delta_{o1} &= \frac{\partial E}{\partial out_{o1}} * \frac{\partial out_{o1}}{\partial net_{o1}} \\ &= -(t\arg et_{o1} - out_{o1}) * out_{o1}(1 - out_{o1}) \end{aligned} \tag{12}$$

So, using Equations (11) and (12) $\frac{\partial E}{\partial w_6}$ becomes, as in (13)

$$\frac{\partial E}{\partial w_6} = \delta_{o1} out_{h1} \tag{13}$$

In order to reduce the error, subtract the obtained value from current weight with η as the learning rate, as in Equation (14)

$$w_6^+ = w_6 - \eta \frac{\partial E}{\partial w_6} \tag{14}$$

At the hidden layer, continue with the backward pass by calculating the new updated values for w_1, w_2, w_3, w_4 and w_5. To do so, use similar procedure as for the output layer but having slight difference since output of every hidden neuron contributes to output and error of the output layer neuron, as in Equation (15).

$$\frac{\partial E}{\partial out_{h1}} = \frac{\partial E}{\partial net_{o1}} * \frac{\partial net_{o1}}{\partial out_{h1}} \tag{15}$$

However, $\frac{\partial E}{\partial net_{o1}}$ is computed, as in Equation (16)

$$\frac{\partial E}{\partial net_{o1}} = \frac{\partial E}{\partial out_{o1}} * \frac{\partial out_{o1}}{\partial net_{o1}} \tag{16}$$

And, using Equation (4) $\frac{\partial net_{o1}}{\partial out_{h1}}$ is obtained, as in Equation (17)

$$\frac{\partial net_{o1}}{\partial out_{h1}} = w_6 \tag{17}$$

Thus, using Equations (15-17) $\frac{\partial E}{\partial out_{h1}}$ becomes, as in Equation (18)

$$\frac{\partial E}{\partial out_{h1}} = \frac{\partial E}{\partial out_{o1}} * \frac{\partial out_{o1}}{\partial net_{o1}} * w_6 \tag{18}$$

Now, using Equation (12) $\frac{\partial E}{\partial out_{h1}}$ becomes, as in Equation (19)

$$\frac{\partial E}{\partial out_{h1}} = \delta_{o1} w_6 \tag{19}$$

Also, it is needed to figure out $\frac{\partial out_{h1}}{\partial net_{h1}}$ using Equation (3) and then $\frac{\partial net_{h1}}{\partial w}$ using Equation (2) for each weight, as in Equations (20) and (21) respectively.

$$\frac{\partial out_{h1}}{\partial net_{h1}} = out_{h1}(1 - out_{h1}) \tag{20}$$

$$\frac{\partial net_{h1}}{\partial w_1} = f_1 \tag{21}$$

Putting it all together, $\frac{\partial E}{\partial w_1}$ is obtained, as in Equation (22)

$$\frac{\partial E}{\partial w_1} = \frac{\partial E}{\partial out_{h1}} * \frac{\partial out_{h1}}{\partial net_{h1}} * \frac{\partial net_{h1}}{\partial w_1} \tag{22}$$

Then using Equations (19-22) $\frac{\partial E}{\partial w_1}$ becomes, as in

Equation (23)

$$\frac{\partial E}{\partial w_1} = \delta_{o1} w_6 * out_{h1}(1 - out_{h1}) * f_1 \quad (23)$$

Now, use the delta δ_{h1}, as in Equation (24)

$$\delta_{h1} = \delta_{o1} w_6 * out_{h1}(1 - out_{h1}) \quad (24)$$

So, using Equations (23) and (24) $\frac{\partial E}{\partial w_1}$ becomes, as in (25)

$$\frac{\partial E}{\partial w_1} = \delta_{h1} f_1 \quad (25)$$

The weight (w_1) is updated, as in Equation (26)

$$w_1^+ = w_1 - \eta \frac{\partial E}{\partial w_1} \quad (26)$$

Rest of the other weights w_2, w_3, w_4 and w_5 can be updated on the same lines. After the first round of backpropagation the total error is only slightly down. After repeating this process 100 times, error plummets to a much larger extent. At that point, the neural network generates the desired output.

4. Experimental Setup, Dataset and Results

This section details the experimental setup, Punjabi dataset and results of the Punjabi summarization task.

4.1. Experimental Setup

For the experimental setup, installation of Python@ version 3.3.7 is quite workable for the Punjabi. Additional python libraries are- NumPy: python numeric, Pandas: analysis of data, lxml Library: web scrapping, pyiwn: Python Package Index (PyPl) API accesses WordNet for the Indian languages-Indo WordNet (here Punjabi language) to extract Punjabi nouns. Unlike the English language which uses the American Standard Code for Information Interchange (ASCII)-American Standard Code for Information Interchange, the Punjabi language is operational with the Unicode. And so, the Punjabi dataset comprises of the encoding-Universal Transformation Format (UTF).

4.2. Punjabi Dataset

The Punjabi dataset is collected as a monolingual Punjabi text corpus under ndian Languages Corpora Initiative (ILCI) phase-II-Indian Languages Corpora Initiative Phase-II. The ILCI project is initiated by the Meity-Ministry of Electronics and Information Technology, Government of India, Jawaharlal Nehru University, New Delhi, India. To access the dataset, researchers can register and login to Technology Development for Indian Languages (TDIL) website [47] which is initiated by the MeitY, from there the Punjabi corpus is freely downloadable [41]. The corpus consists of 30,000 Punjabi sentences from general domain. The corpus based Punjabi sentences are Part-Of-Speech (POS) tagged, as per the Bureau of Indian Standards (BIS) tagset which ensures adequate representation of the language within language technology standards.

Figure 2 shows that each of the Punjabi sentences has UTF encoding in a text file format. Since some researchers are unaware of the Punjabi language. For this purpose, in this paper, the Punjabi dataset is made understandable by them while looking at the transliteration and English translation of the Punjabi sentences as in Figure 3.

ਇਹ ਪ੍ਰਾਂਤ ਹਰ ਉਮਰ ਦੇ ਲੋਕਾਂ ਦੀਆਂ ਜ਼ਰੂਰਤਾਂ ਪੂਰੀਆਂ ਕਰਦਾ ਹੈ।

...

ਇਸ ਸੈਂਟਰ ਦਾ ਮਕਸਦ ਬੁੱਧ ਦੇ ਸਿਧਾਂਤਾਂ ਨਾਲ ਮਨੁੱਖ ਨੂੰ ਜੋੜਨਾ ਹੈ।

...

ਹੇਮਕੁੰਟ ਸਾਹਿਬ ਅਤੇ ਇਸ ਨਾਲ ਸੰਬੰਧਿਤ ਅਸਥਾਨਾਂ ਅਤੇ ਗੁਰਦੁਆਰਿਆਂ ਦਾ ਪ੍ਰਬੰਧ ਟਰੱਸਟ ਦੁਆਰਾ ਚਲਾਇਆ ਜਾਂਦਾ ਹੈ।

Figure 2. Sample Punjabi dataset.

Transliteration: Iha prânta hara umara dē lōkāṁ dī'āṁ zarūratāṁ pūrī'āṁ karadā hai.
English Translation: This province meets the needs of people of all ages.
...

Transliteration: Isa saiṇṭara dā makasada budha dē sidhāntāṁ nāla manukha nū jōṛanā hai.
English Translation: The purpose of this center is to connect man with the principles of Buddhism.
...

Transliteration: Hēmakuṇa sāhiba atē isa nāla sabadhita asathānāṁ atē guradu'ārī'āṁ dā prabadha ṭarasaṭa du'ārā calā'i'ā jāndā hai.
English Translation: Hemkunt Sahib and its associated places and gurdwaras are administered by the Trust.

Figure 3. Transliteration and translation of sample Punjabi dataset.

4.3. Results

In order to better interpret the summarized results, consider a Punjabi sentence (Figure 4) that is included in the generated summary by the NN system.

ਕਿਸਾਨਾਂ ਦੀ ਆਮਦਨ ਦੁੱਗਣੀ ਕਰਨ ਨਾਲ ਜੀਐਸਟੀ ਪ੍ਰਭਾਵਤ ਹੋਏਗਾ।
Transliteration: Kisānāṁ dī āmadana dugaṇī karana nāla jī'aisaṭī prabhāvata hō'ēgā.
English Translation: Doubling the income of farmers will effect the GST.

Figure 4. A sample Punjabi sentence.

On encompassing the pre-processing phase, the considered Punjabi sentence is cleaned as in Figure 5.

ਕਿਸਾਨ ਆਮਦਨ ਦੁੱਗਣੀ ਜੀਐਸਟੀ ਪ੍ਰਭਾਵਤ
Transliteration: Kisāna āmadana dugaṇī jī'aisaṭī prabhāvata
English Translation: Farmer income doubles GST affected

Figure 5. Pre-processed Punjabi sentence.

The processing phase extracts feature values, as in Table 4 i.e., TF-ISF (f_1): 0.36, headline and next line (f_2): 0, NER (f_3): 1, cue-phrase (f_4): 0, and CPEN (f_5): 1.

Table 4. Feature values for example sentence.

Feature	Value	Remarks
TF-ISF (f_1)	0.36	1/5 (log(60))
Headline and next line (f_2)	0	doesn't constitute headline
NER (f_3)	1	ਕਿਸਾਨ *kisāna 'farmer'*
Cue-phrase (f_4)	0	no cue within sentence
CPEN (f_5)	1	ਜੀਐਸਟੀ *jī'aisaṭī 'GST'*

The neural network phase initializes the random weights with respect to each feature as in Table 5, so feature-weight combination becomes: (f_1, w_1) as (0.36, 0.20); (f_2, w_2) as (0, 0.30); (f_3, w_3) as (1, 0.40); (f_4, w_4) as (0, 0.50); and (f_5, w_5) as (1, 0.60) respectively.

Table 5. Weighted feature values for example sentence.

Feature and Associated Weight	Initial Weight	Weighted Feature Value
f_1, w_1	0.20	0.0720
f_2, w_2	0.30	0.0000
f_3, w_3	0.40	0.4000
f_4, w_4	0.50	0.0000
f_5, w_5	0.60	0.6000

The processing of forward pass within the backpropagation neural network is seen in Table 6. To compute net_{o1}, the weighted connection w_6 is 0.35.

Table 6. Forward pass for example sentence.

net_{h1}	out_{h1}	net_{o1}	out_{o1}
2.07200	0.607429410150	1.21245	0.444329277680

The error within the generated output is computed with respect to the target (0.45) threshold as is seen in Table 7. The computed error is quite less which indicates that the weights are approachable to convergence. From the backward pass, the updated weight w_6 (w_6^+) is obtained.

Table 7. Error computation and weight updation.

Error	Updated (w_6)
0.00283536115	0.28905920013

The rest other weights are then updated too, and their updated values are at par as seen in Table 8. Thus, the chosen sentence is included in the generated summary.

Table 8. Updated weights in backward pass.

Updated Weight	Weight Value
w_1^+	0.20005849783
w_2^+	0.30005849783
w_3^+	0.40005849781
w_4^+	0.5000584973
w_5^+	0.6000584873

Overall, the highest scoring sentences are picked up and are added to the summary file (Summary.txt). However, sentences as in Figure 6 are discarded off as they do not pass the classification phase.

ਕੀ ਇਹ ਆਮ ਆਦਮੀ 'ਤੇ ਅਸਰ ਨਹੀਂ ਪਾਏਗੀ?
Transliteration: Kī iha āma ādamī'tē asara nahīṁ pā'ēgī?
English Translation: Will it not affect the common man?

Figure 6. Punjabi sentence excluded from summary.

As a measure of the summary evaluation, F-measure, as in Equation (27) is computed for the Punjabi dataset. The dataset considers one human reference summary, and evaluation of a system generated summary is compared with the reference summary.

$$\text{F-measure} = \frac{(2 * Precision * Recall)}{(Precision + Recall)} \quad (27)$$

Where, Precision and Recall are defined as in Equations (28) and (29) respectively.

$$\text{Precision} = \frac{Number\ of\ correct\ sentences\ retrieved\ by\ system}{Total\ number\ of\ sentences\ retrieved\ by\ system} \quad (28)$$

$$\text{Recall} = \frac{Number\ of\ correct\ sentences retrieved\ by\ system}{Total\ number\ of\ sentences retrieved\ by\ human} \quad (29)$$

The neural networks based Punjabi text summarizer has achieved the precision as 90.02%, recall as 89.28%, and F-measure as 89.65% respectively. The performance of the NN based Punjabi summarizer is quite reasonably good as compared to the performance of other existing Indian languages summarizers (Table 9).

Table 9. Comparison of Indian languages summarizers.

Summarization System	Dataset	Methodology	Evaluation
Proposed Punjabi Summarization System	Indian Languages Corpora Initiative Phase-II	Neural Networks	Precision: 90.02% Recall: 89.28% F-measure: 89.65%
Bengali Summarizer [42]	38 Bengali documents from Bengali newspaper-Ananda Bazar Patrika	TF-IDF	Recall: 41.22%
Multi-Lingual Summarizer (English, Hindi, Gujarati and Urdu) [39]	English: DUC 2002 Hindi: Dainik Jagaran Gujarati: Bhashaindia Urdu: BBC News	Fuzzy Logic	Accuracy: 82%
Kannada Summarizer [25]	Document Categories- Literature Entertainment Astrology Sports	GSS coefficients, TF-IDF	Recall: Literature: 70% Entertainment: 80% Astrology: 80% Sports: 76%
Hindi Summarizer [34]	Documents from different domains	Graph based	Precision: 79% Recall: 69% F-measure: 70%

5. Conclusions and Future Work

This paper proposes a three phase Punjabi text summarization methodology which undergoes the preprocessing, processing and classification phases respectively. For this, monolingual Punjabi text corpus from Indian languages corpora initiative phase-2 is taken which consists of 30,000 Punjabi sentences in

the UTF-encoding. The preprocessing phase cleans the Punjabi text; processing phase extracts the statistical and linguistic features; and classification based neural network undergoes weights inclusion to features during the forward pass and weights updation during the backward pass until either they converge or suitable number of iterations is accomplished. Punjabi sentences that clearly pass the neural network-based backpropagation are exemplified. As a result, the highest scored Punjabi sentences are added into the generated summary. Then the proposed Punjabi text summarizer has achieved precision (90.02%), recall (89.28%), and F-measure (89.65%) respectively.

In future, the summarization methodology can be compared with other classification techniques such as support vector machines and many more while incorporating the labeled Punjabi data. Also, one can add more features with profound understandability of the Punjabi text in order to produce abstractive summaries. And, the summary system can be made language and platform independent too.

Acknowledgement

This research is partially funded by the Ministry of Economy, Industry and Competitiveness, Spain (CSO2017-86747-R).

References

[1] Al-Abdallah R. and Al-Taani A., "Arabic Single-Document Text Summarization Using Particle Swarm Optimization Algorithm," *Procedia Computer Science*, vol. 117, pp. 30-37, 2017.

[2] Aries A., Zegour D., and Hidouci W., "Automatic Text Summarization: What Has Been Done and What Has to Be Done," *arXiv preprint arXiv:1904.00688*, 2019.

[3] Aslam J., Diaz F., Ekstrand-Abueg M., McCreadie R., Pavlu V., and Sakai T., "TREC 2015 Temporal Summarization Track Overview," *in Proceedings of National Institute of Standards and Technology*, Gaithersburg MD, 2015.

[4] Dalal V. and Malik L., "Automatic Summarization for Hindi Text Documents Using Bio-Inspired Computing," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 6, no. 4, pp. 682-688, 2017.

[5] Dalal V. and Malik L., "Semantic Graph based Automatic Text Summarization for Hindi Documents Using Particle Swarm Optimization," *in Proceedings of the International Conference on Information and Communication Technology for Intelligent Systems*, Ahmedabad, pp. 284-289, 2017.

[6] Gill M., Lehal G., and Joshi S., "Part of Speech Tagging for Grammar Checking of Punjabi," *The Linguistic Journal*, vol. 4, no. 1, pp. 6-21, 2009.

[7] Gulati A. and Sawarkar S., "A Novel Technique for Multi-Document Hindi Text Summarization," *in Proceedings of International Conference on Nascent Technologies in Engineering*, pp. 1-6, Vashi, 2017.

[8] Gupta V., "Hybrid Algorithm for Multilingual Summarization of Hindi and Punjabi Documents," *in Proceedings of Mining Intelligence and Knowledge Exploration*, Tamil Nadu, pp. 717-727, 2013.

[9] Gupta V. and Kaur N., "A Novel Hybrid Text Summarization System for Punjabi Text," *Cognitive Computation*, vol. 8, no. 2, pp. 261-277, 2016.

[10] Gupta V. and Lehal G., "A Survey of Text Summarization Extractive Techniques," *Journal of Emerging Technologies in Web Intelligence*, vol. 2, no. 3, pp. 258-268, 2010.

[11] Gupta V. and Lehal G., "Automatic Keywords Extraction for Punjabi Language," *International Journal of Computer Science Issues*, vol. 8, no. 5, pp. 327-331, 2011.

[12] Gupta V., "Automatic Normalization of Punjabi Words," *International Journal of Engineering Trends and Technology (IJETT)*, vol. 6, no. 7, pp. 353-357, 2013.

[13] Gupta V. and Lehal G., "Automatic Text Summarization System for Punjabi Language," *Journal of Emerging Technologies in Web Intelligence*, vol. 5, no. 3, pp. 257-271, 2013.

[14] Gupta V. and Lehal G., "Complete Pre-Processing Phase of Punjabi Text Extractive Summarization System," *in Proceedings of the COLING 2012: Demonstration Papers*, Mumbai, pp. 199-206, 2012.

[15] Gupta V. and Lehal G., "Named Entity Recognition for Punjabi Language Text Summarization," *International Journal of Computer Applications*, vol. 33, no. 3, pp. 28-32, 2011.

[16] Gupta V. and Lehal G., "Preprocessing Phase of Punjabi Language Text Summarization," *in Proceedings of the International Conference on Information Systems for Indian Languages*, Patiala, pp. 250-253, 2011.

[17] Houtinezhad M. and Ghaffary H., "Improvement Of Imperialist Competitive Algorithm Based on the Cosine Similarity Criterion of Neighboring Objects," *The International Arab Journal of Information Technology*, vol. 18, no. 3, pp. 261-269, 2021.

[18] Hu H., Tang L., Zhang S., and Wang H., "Predicting the Direction of Stock Markets Using Optimized Neural Networks with Google Trends," *Neurocomputing*, vol. 285, pp. 188-195, 2018.

[19] Jain A., Named Entity Recognition for Hindi Language Using NLP Techniques, PhD Thesis, Jaypee Institute of Information Technology, Noida, 2019.

[20] Jain A. and Arora A., "Named Entity Recognition in Hindi Using Hyperspace Analogue to Language and Conditional Random Field," *Pertanika Journal of Science and Technology*, vol. 26, no. 4, pp. 1801-1822, 2018.

[21] Jain A. and Arora A., "Named Entity System for Tweets in Hindi Language," *International Journal of Intelligent Information Technology*, vol. 14, no. 4, pp. 55-76, 2018.

[22] Jain A., Tayal D., Yadav D., and Arora A., *Data Visualization and Knowledge Engineering*, Springer, 2020.

[23] Jain A., Yadav D., and Arora A., "Particle Swarm Optimization for Punjabi Text Summarization," *International Journal of Operations Research and Information Systems*, vol. 12, no. 3, pp. 1-17, 2021.

[24] Jain A., Yadav D., and Tayal D., "NER for Hindi Language Using Association Rules," *in Proceedings of the International Conference on Data Mining and Intelligent Computing*, New Delh, pp. 1-5, 2014.

[25] Jayashree R., Srikanta K., and Sunny K., "Document Summarization in Kannada Using Keyword Extraction," *in Proceedings of the American Institute of Aeronautics and Astronautics AIAA*, pp. 121-127, 2011.

[26] Kanapala A., Pal S., and Pamula R., "Text Summarization from Legal Documents: A Survey," *Artificial Intelligence Review*, vol. 51, no. 1, pp. 371-402, 2019.

[27] Kaur A., Josan G., and Kaur J., "Named Entity Recognition for Punjabi: A Conditional Random Field Approach," *in Proceedings of the 7th International Conference on Natural Language Processing*, India, 2009.

[28] Kaur A., Singh P., and Kaur K., "Punjabi Dialects Conversion System for Majhi, Malwai and Doabi Dialects," *in Proceedings of the 8th International Conference on Computer Modelling and Simulation*, New York, pp. 125-128, 2017.

[29] Kaur J. and Saini J., "Punjabi Stop Words: A Gurmukhi, Shahmukhi and Roman Scripted Chronicle," *in Proceedings of the ACM Symposium on Women in Research*, New York, pp. 32-37, 2016.

[30] Kaur R. and Sharma S., "Semi-Automatic Domain Ontology Graph Generation System in Punjabi," *in Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*, New York, pp. 1-5, 2016.

[31] Kaur R., Sharma R., Preet S., and Bhatia P., "Punjabi Wordnet Relations and Categorization of Synsets," *in Proceedings of the 3rd National Workshop on IndoWordNet Under the Aegis of the 8th International Conference on Natural Language Processing*, Kharagpur, 2010.

[32] Krail N. and Gupta V., "Domain Based Classification of Punjabi Text Documents Using Ontology and Hybrid-Based Approach," *in Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language* Processing, Mumbai, pp. 109-122, 2012.

[33] Kumar K. and Yadav D., "An Improvised Extractive Approach to Hindi Text Summarization," *in Proceedings of Information Systems Design and Intelligent Applications*, India, pp. 291-300, 2015.

[34] Kumar K., Yadav D., and Sharma A., "Graph Based Technique for Hindi Text Summarization," *in Information Systems Design and Intelligent Applications*, India, pp. 301-310, 2015.

[35] Lin J., Roegiest A., Tan L., McCreadie R., Voorhees E., and Diaz F., "Overview of the TREC 2016 Real-Time Summarization," *in Proceedings of the Text REtrieval Conference TREC*, Gaithersburg, pp. 15-18, 2016.

[36] Liu Y. and Lapata M., "Text Summarization with Pretrained Encoders," *in Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, pp. 3728-3738, 2019.

[37] Mohd M., Jan R., and Shah M., "Text Document Summarization Using Word Embedding," *Expert Systems with Applications*, vol. 143, pp. 112958, 2020.

[38] Nasser I. and Abu-Naser S., "Lung Cancer Detection Using Artificial Neural Network," *International Journal of Engineering and Information Systems*, vol. 3, no. 3, pp. 17-23, 2019.

[39] Patel A., Siddiqui T., and Tiwary U., "A Language Independent Approach to Multilingual Text Summarization," *in Proceedings of the Conference RIAO2007, Pittsburgh PA*, Paris, pp. 123-132, 2007.

[40] Prudhvi K., Chowdary A., Reddy P., and Prasanna P., "Text Summarization Using Natural Language Processing," *in Proceedings of Intelligent System Design*, pp. 535-547, 2021.

[41] Punjabi Monolingual Text Corpus: *http://www.tdil-dc.in/index.php?option=com_download&task=showresourceDetails&toolid=1890&lang=en*, Last Visited, 2020.

[42] Sarkar K., "Bengali Text Summarization by Sentence Extraction," *in Proceedings of the*

International Conference on Business and Information Management, pp. 233-245, 2012.

[43] See A., Liu P., and Manning C., "Get To The Point: Summarization with Pointer-Generator Networks," *arXiv preprint arXiv:1704.04368*, 2017.

[44] Singh A. and Singh P., "Punjabi Dialects Conversion System for Malwai and Doabi Dialects," *Indian Journal of Science and Technology*, vol. 8, no. 27, pp. 1-7, 2015.

[45] Singh S., Kumar A., Mangal A., and Singhal S., "Bilingual Automatic Text Summarization Using Unsupervised Deep Learning," *in Proceedings of International Conference on Electrical, Electronics, and Optimization Techniques*, Chennai, pp. 1195-1200, 2016.

[46] Subramaniam M. and Dalal V., "Test Model For Rich Semantic Graph Representation for Hindi Text Using Abstractive Method," *International Research Journal of Engineering and Technology*, vol. 2, no. 2, pp. 113-116, 2015.

[47] TDIL: http://www.tdil-dc.in/index.php?lang=en, last visited, 2020.

[48] Tsenov G. and Mladenov V., "Speech Recognition Using Neural Networks," *in Proceedings of 10th Symposium on Neural Network Applications in Electrical Engineering*, Belgrade, pp. 181-186, 2010.

[49] Wyllys R., "Extracting and Abstracting by Computer," *Automated Language Processing*, pp. 127-179, 1967.

[50] Xu H., Li K., Wang Y., Wang J., Kang S., Chen X., Povey D., and Khudanpur S., "Neural Network Language Modelling With Letter-Based Features and Importance Sampling," *in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, pp. 6109-6113, 2018.

[51] Yong S., Abidin A., and Chen Y., "A Neural-Based Text Summarization System," *WIT Transactions on Information and Communication Technologies*, vol. 37, pp. 185-192, 2006.

Arti Jain has received her Ph.D, in Computer Science and Engineering (2019) from Jaypee Institute of Information Technology, Noida, India. She has more than 18 years of teaching experience. Currently, she is working as Assistant Professor (Sr. Grade), CSE, JIIT, Noida, India. She is a member of IEEE, INSTICC, IAENG, IFERP and Life Member of TERA. She is reviewer of several reputed International Journals and TPC member of International Conferences. She has more than 20 publications in peer-reviewed International Journals, Book Chapters and International Conferences. Her research interests include Natural Language Processing, Machine Learning, Data Science, Deep Learning, Social Media Analysis, Big Data and Data Mining.

Anuja Arora is working as Associate Professor in the Department of Computer Science and Engineering at Jaypee Institute of Information Technology, Noida, India. She is having academic experience of 15 years and industry experience of 1.5 years. She is Senior IEEE Member, ACM Member, SIAM Member, INSTICC and Life Member of IAENG. She is also Vice-Chair for the Delhi ACM-W Chapter. She has more than 70 research papers in peer-reviewed International Journals, Book Chapters, and International Conferences. She has supervised 3 Ph.D. thesis and 2 more are in progress. Her research interest includes Data Science, Deep Learning, Information Retrieval Systems, Machine Learning, Social Network Analysis, Software Testing and Web Intelligence. She is reviewer of many reputed and peer-reviewed IEEE transactions- TKDE, TNSM, IEEE Transaction of Cybernetics, Springer, IGI Global, Inderscience, and De Gruyter Journals. She has guided more than 17 M.Tech Thesis and around 100 B.Tech Projects.

Divakar Yadav (SM'2017) is working as Associate Professor in the Department of Computer Science and Engineering at National Institute of Technology (NIT), Hamirpur (HP), India. He did his undergraduate in Computer Science and Engineering (1999), Post Graduate in Information Technology (2005) and PhD in Computer Science and Engineering (2010). He is Senior Member IEEE. He has also worked as Post-Doctoral Fellow at University of Carlos-III, Madrid, Spain from 2011-2012. He has supervised 5 PhD thesis and 22 Master dissertations. He has more than 20 years of teaching and research experience. He has published 85 research articles in reputed International Journals and Conference Proceedings. His area of research is Machine Learning and Information Retrieval.

Jorge Morato has received his B.S. degree in Sciences from University of Alcala in 1992, and Ph.D. degree in Information Sciences from University Carlos III in 1999. Since 2000, he is researcher and professor with the Computer Science Department, Carlos III University, Spain. From 1991-2016, he has had grants and fellowships from the Spanish National Research Council and the Spanish Government. He is the author of more than one hundred papers and book chapters. His research interests include NLP applications, Information Retrieval Algorithms, Web Positioning and Readability, and Knowledge Organization Systems.

Amanpreet Kaur has received her B.E. degree in CSE from Guru Nanak Dev University, Amritsar, India, 2006, M.E. degree in Computer Science Engineering from NIT Jalandhar, 2009, and Ph.D. in Computer Science and Engineering from Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India, 2020. She is currently working as Assistant Professor, CSE, Jaypee Institute of Information Technology, Noida (UP), India. Her research interests include Wireless Sensor Networks, Information Security and Performance Analysis.