International Journal of Operations Research and Information Systems
Volume 12 • Issue 3 • July-September 2021

Particle Swarm Optimization for Punjabi Text Summarization

Arti Jain, Jaypee Institute of Information Technology, Noida, India
https://orcid.org/0000-0002-3764-8834

Divakar Yadav, National Institute of Technology, Hamirpur, India
https://orcid.org/0000-0001-6051-479X

Anuja Arora, Jaypee Institute of Information Technology, Noida, India
https://orcid.org/0000-0001-5215-1300

ABSTRACT

Particle swarm optimization (PSO) algorithm is proposed to deal with text summarization for the Punjabi language. PSO is based on intelligence that predicts among a given set of solutions which is the best solution. The search is carried out by extremely high-speed particles. It updates particle position and velocity at the end of iteration so that during the development of generations, the personal best solution and global best solution are updated. Calculation within PSO is performed using fitness function which looks into various statistical and linguistic features of the Punjabi datasets. Two Punjabi datasets—monolingual Punjabi corpus from Indian Languages Corpora Initiative Phase-II and Punjabi-Hindi parallel corpus—are considered. The parallel corpus comprises 1,000 Punjabi sentences from the tourism domain while monolingual corpus contains 30,000 Punjabi sentences of the general domain. ROUGE measures evaluate summary where the highest measure, ROUGE-1, is achieved for parallel corpus with precision, recall, and F-measure as 0.7836, 0.7957, and 0.7896, respectively.

KEYWORDS

Particle Position, Particle Swarm Optimization, Particle Velocity, Punjabi Dataset, ROUGE Measure, Text Summarization

1. INTRODUCTION

In the modern era, utilization of digital content has raised dramatically at all walks of our life, for example- social media: facebook (Jain et al., 2018a), twitter (Jain et al., 2018b); newswire articles (Jain et al., 2014), web corpora: health corpus (Jain et al., 2018c), tweet corpus (Jain and Arora, 2018); web advertisements (Jain et al., 2013), question answering system (Verma et al., 2020) and many more. The dire consequence is a heavy flood of information over the internet. Thus, lays a necessity to build an automated system to summarize the text that provides a meaningful and concise summary.

DOI: 10.4018/IJORIS.20210701.oa1

Earlier, the text summarization task primarily involved statistical features (Meena and Gopalani, 2016) to work upon. Later on, morphological analysis (Almazaydeh, 2018) is also done due to availability of new technologies and labeled data. Now-a-days, the text summarization process has reached to a mature stage for English (Saggion and Poibeau, 2013; Gambhir and Gupta, 2017) and other foreign languages (Gunawan et al., 2017). However, for Indian languages the text summarization is a way behind as several language based challenges persist that are unfold in this paper. Our aim is to carry forward the Text Summarization (TS) task for the Punjabi language (Gupta and Kaur, 2016) which is still in its premature stage because of scarcity of the labeled data and other constraints.

Punjabi is the third most spoken language in the Indian subcontinent with more than 100 million native speakers around the world. It is an official language of the Punjab state which encompasses northwest India and eastern Pakistan. It is the most widely spoken language in Pakistan, second/third language used by around 30 million people in India. In addition, Punjabi is a minority language in several other countries where Punjabi people have migrated, namely- United States of America, Australia, United Kingdom, and Canada. The Punjabi language comprises of canonical word order of Subject Object Verb (SOV), also contains postpositions; distinguishes gender- masculine/feminine, number- singular/plural, case- direct/oblique. The major writing system is the ਗੁਰਮੁਖੀ *'Gurmukhi'*, a Punjabi script. However, very fewer efforts are being made in the field of computer technology towards the development of the enriched Punjabi language.

Text summarization (Al-Zahrani et al., 2015) is one of the vital Natural Language Processing (NLP) (Jain, 2019) tasks that mainly consists of two phases- pre-processing and processing phases. In the pre-processing phase, various keywords & clauses are taken care using linguistic and statistical features. In the processing phase, two techniques- extractive and abstractive summarization are used. Extractive summarization (Lins et al., 2020) is a shallow technique where an extracted text is used as a summary and is comparatively easier to implement. Abstractive summarization (Song et al., 2019) requires deep understanding and analysis of a given text which is comparatively more complex to implement. In the abstractive summary, new concepts and expressions determine the original text in a concise form to convey relevant information, henceforth, summary may not contain the identical sentences as in the original document.

In this paper, text summarization task in Punjabi using Particle Swarm Optimization (PSO) (Al-Abdallah and Al-Taani, 2017) algorithm is proposed which induces a meaningful short summary out of a large web text- Unicode encoded Punjabi text. PSO is one of the most powerful bio-inspired optimization techniques which looks into the bird-flock/fish-school concept that is precise and easier to implement. The PSO is based upon intelligence which has neither overlapping nor mutation calculation issues as in the Genetic Algorithm (GA). PSO helps to predict that among a given set of solutions which one is the best solution. The search is carried out by extremely high-speed particles. It updates particle position and velocity at the end of iteration, so that during the development of generations, the personal best solution of the particle and the global best solution are updated. Calculation within PSO is quite simple as it contains greater optimal capability which can be fulfilled with ease. PSO adapts the real number code which is determined by the solution itself. Thus, PSO based text summarization is quite suitable to extract summary sentences from an input Punjabi text document, grades the sentences, and actual summary is executed as the topmost summary.

The rest of the paper is organized into the following sections. Section 2 explains the literature relevant to PSO algorithm and Punjabi language processing tasks. Section 3 illustrates existing tools and libraries for the Gurmukhi script. Section 4 details the proposed PSO based TS methodology for the Punjabi text. Section 5 describes experimental setup, Punjabi datasets and results. At the end, Section 6 is the conclusion of the paper.

2. RELATED WORK

The related summarization task based on PSO algorithm is mentioned below.

(Binwahlan et al., 2009) have discussed feature selection strategy where 5 features relevant to summarization are used; and PSO is worked with to learn and obtain the weights of every feature. The PSO based summary is forty-three percent similar to the summary that is generated by the human. (Abuobieda et al., 2011) have illustrated a feature selection-based PGP Sum - pseudo genetic probabilistic summarization model which generates single document based extractive summary. PGP Sum obtains the weights of features which tunes the features scores and optimize summarization task. PGP Sum model similarity ratio is quite closer to the summary that is generated by the human. (Al-Abdallah and Al-Taani, 2017) have compared PSO with Genetic Algorithm (GA) and Harmony Search (HS) for Arabic text summarization over Essex Arabic Summaries Corpus (EASC). It is observed that PSO outperforms GA and HS over ROUGE scores. (Zhang et al., 2018) have proposed Adaptive Particle Swarm Optimization (AdPSO) algorithm which is based on an aggregation degree. It prevents PSO from falling into the local optimum, and from low convergence precision. It improves the global search capability while introducing random disturbances to the search space via Differential Evolution (DE). It improves various factors- diversity of population, activity of particles, performance of convergence, and ability to search within PSO. (Zhang et al., 2020) have proposed Multi-Objective Particle Swarm Optimization based Community Discovery (MOPSO-CD) algorithm which thwart from fall in the local optimal solution. It has higher effectiveness and better community detection capability as it adjusts the population to improve accuracy which is caused by the uncertainty of PSO.

The related language processing tasks for Punjabi are as follows: (Kaur et al., 2010) have discussed assessment techniques for the Punjabi WordNet which has considered relations and categorization of synsets among the Punjabi words. (Sharma and Jhajj, 2010) have detailed optical character recognition in Punjabi language using zoning for feature extraction and classification techniques- K-Nearest Neighbor (KNN) and Support Vector Machines (SVM). (Gupta and Lehal, 2011a) have dealt Named Entity Recognition (NER) task for Punjabi text to identify entities such as name of person, location and organizations using rule-based and list lookup approach. (Gupta and Lehal, 2011b) have detailed an automatic keywords extraction task which has identified key phrases/keywords/key segments from the Punjabi text. The task includes phases such as removal of stop-words, nouns identification and stemming, computation of Term Frequency and Inverse Sentence Frequency (TF-ISF). More than 50 Punjabi documents from the Punjabi news corpus are taken which gives precision of 80.40%, recall of 90.60% and F-measure 85.20% respectively. (Kaur and Gupta, 2011) have implemented topic tracking using approaches- Vector Space Model (VSM), KNN classification, hierarchical clustering and others. (Nidhi, 2012) has worked upon the ontological concepts and hybrid algorithm for the Punjabi document's classification task. (Kaur and Kaur, 2013) have presented deadwood detection and elimination method to eliminate unwanted/unnecessary words or phrases which carried no meaning to the sentences. (Kaur and Gupta, 2014) have discussed the concept of sentiment analysis in the Punjabi while using lexicon, unigram and simple scoring method. (Singh and Kaur, 2014) have presented a simple and easy to implement rule-based approach for shallow question generation system using NER tool. The tool has recognized names from the Punjabi text containing historical information, and generated appropriate questions from them. (Singh and Singh, 2015) have discussed translation of Punjabi dialects- Malwai and Doabi which are generated by pronunciation, and is achieved by bilingual wordbooks along with morphological rules. (Gupta and Kaur, 2016) have proposed the summarization strategy based on hybrid concepts in the Punjabi language. Their system implements SVM classifier for distinguished features- concept based, statistics based, location based and linguistic based in the Punjabi and compared with ten baseline systems over one-hundred and fifty randomly selected documents from two Punjabi datasets. (Sarkar et al., 2016) have worked upon removal of sentence ambiguity by detection of paraphrases within the Punjabi language while using three similarity measures and training the probabilistic neural networks using feature set. (Sharma, 2017) has mentioned separation of dependent and independent clauses from a sentence in the Punjabi language using some rules and Part of Speech (POS) (Gupta et al., 2011) tagging. (Singh et al., 2018) have elaborated suicide case of farmers in Punjab which is analyzed using morphological and

Figure 1. Punjabi POS Tagger

sentiment analysis in the Gurmukhi script; later on, deep neural network is trained using feature set and splitter to split training and test sets. (Sharma, 2019) has developed syntactic analysis system in the Punjabi for language-based compound sentences. (Ahmad et al., 2020) have developed first-ever NER corpus for the Shahmukhi, another Punjabi script. The corpus comprises of 318,275 tokens and 16,300 named entities. They have compared the corpus specifications with the Gurmukhi counterpart using machine learning and deep learning techniques.

3. EXISTING TOOLS AND LIBRARIES FOR GURMUKHI SCRIPT

Some of the commonly available tools and libraries for the Gurmukhi script are described here.

3.1 Punjabi POS Tagger

POS tagger for Punjabi language (Punjabi POS Tagger)[1] is developed at Punjab University, Patiala, India. Figure 1 inputs the Punjabi text e.g. ਸ਼ੇਰ *Śēra lion'* and obtains the tagged output as ਸ਼ੇਰ\N_NN. It indicates that the inputted text has POS tag as N_NN (Common Noun).

3.2 IndoWordNet

IndoWordNet[2] is developed at the Center for Indian Languages Technology (CFILT), IIT Bombay, India which supports number of Indian languages including Punjabi. Figure 2 depicts the noun morph and synset details for the Punjabi text e.g. ਖੇਤ *khēta 'farm'*.

3.3 Punjabi WordNet

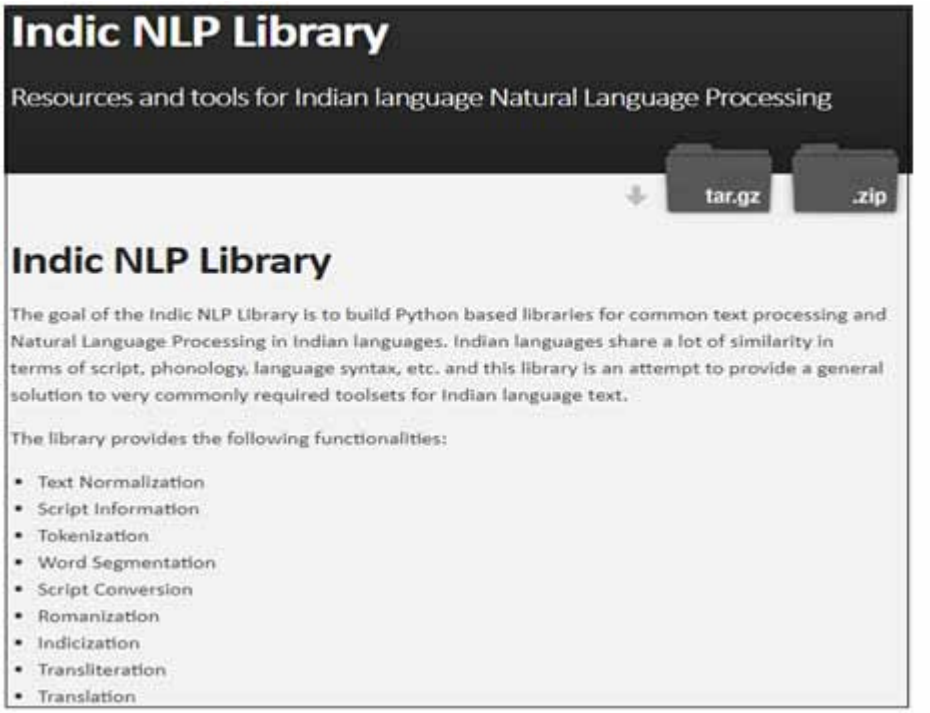
Punjabi WordNet (Narang et al., 2013) is developed by the Thapar University, Patiala, India which provides relevant information of Punjabi words such as word category, synset, concept and example sentences. Figure 3 gives details of the Punjabi text e.g. ਘਰ *ghara 'home'*.

Figure 2. IndoWordNet for Punjabi Text

3.4 Indic NLP Library

Figure 3. Punjabi WordNet

Figure 4. Indic NLP Library



Indic NLP (Indic NLP Library)[3] is a python library that is available on the GitHub which provides various libraries to tokenize sentences/words, normalization of words etc. as is seen in Figure 4. For example, normalizer for the Gurumukhi script is illustrated in Figure 5.

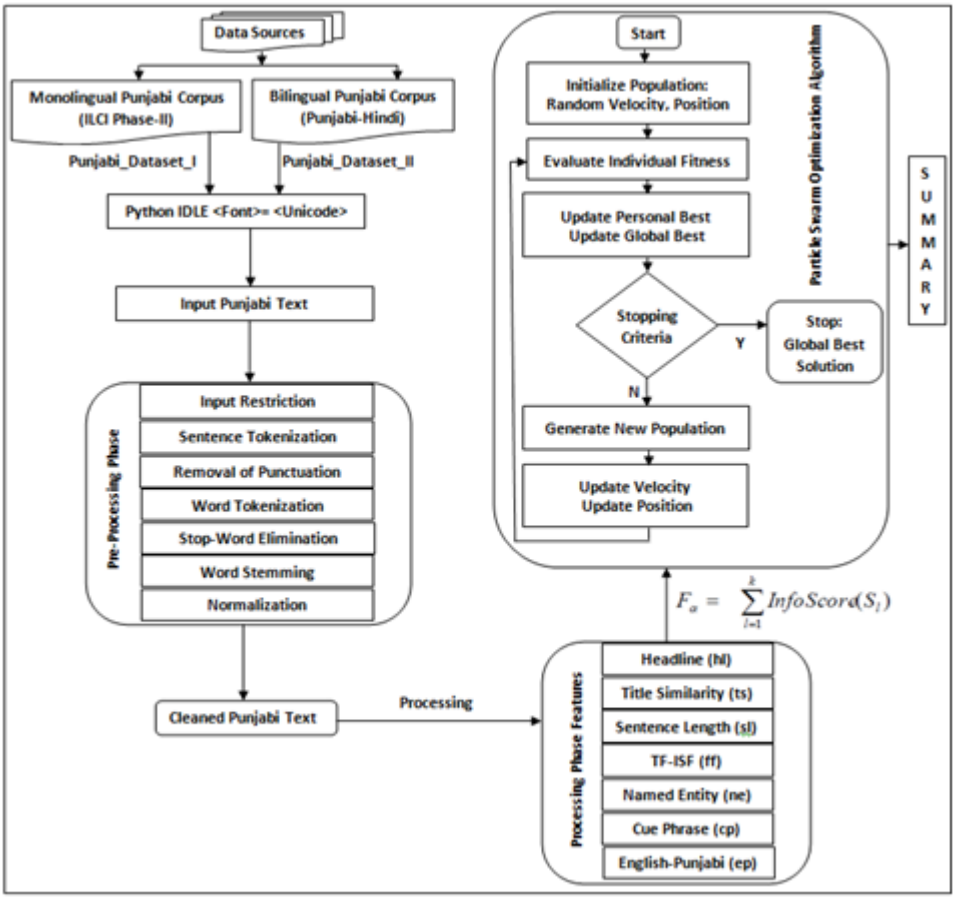
4. PROPOSED FRAMEWORK FOR SUMMARIZATION OF PUNJABI TEXT

The proposed framework for the summarization task of the Punjabi text is detailed in Figure 6. Punjabi text is mainly extracted from two web sources (Section 5.2) - monolingual Punjabi corpus and bilingual Punjabi-Hindi corpus which are converted into Unicode format with the help of python libraries (Section 5.1). Then the Punjabi Unicode text undergoes the pre-processing & processing phases respectively. The first phase- pre-processing accompanies various operations (Section 4.1) - input restriction, sentence tokenization, removal of punctuation, word tokenization, stop-word elimination, word stemming and normalization. Resultant cleaned Punjabi text undergoes the processing phase (Section 4.2) where several features such as headline, title summary, sentence length, TF-ISF, NER (Jain et al., 2020), cue phrase, and English-Punjabi common nouns are applied and sentence scores are calculated. Weight is assigned to each feature, and sentence informative score is calculated to generate fitness function for the particle swarm optimization algorithm. PSO undergoes the following steps (Section 4.3) - initialization of population, current best solution, update particle velocity and

Figure 5. Normalizer for Gurmukhi Script

```
Class: indicnlp.normalize.indic_normalize.GurmukhiNormalizer (remove_nuktas=False)
Bases: indicnlp.normalize.indic_normalize.NormalizerI
```

Figure 6. Proposed Framework



position, and stopping criteria. Once the stopping criteria is met, the Punjabi text summary is generated which is evaluated over ROUGE measures (Section 5.3).

The text summarization methodology based on PSO for the Punjabi text is split into pre-processing phase and processing phase respectively.

4.1 Pre-Processing Phase

In this phase, data cleaning on Punjabi text is achieved using the following major operations- input restriction, sentence tokenization, removal of punctuation, word tokenization, stop-word elimination, word stemming, and normalization. Each of these is discussed in this section.

4.1.1 Input Restriction

Text corpus must be majorly in the Punjabi language which comprises of total number of Punjabi characters as not lesser than 80% of the total number of corpus characters.

4.1.2 Sentence Tokenization

Existence of certain symbols such as ";", "।", "?", "!" are indicators of sentence boundary, particularly, the vertical bar "।" is used for the completion of the Punjabi sentence.

Table 1. Punjabi stop-word, transliteration and English translation

Punjabi Word	Transliteration	English Translation
ਦੇ	dē	of
ਵਚਿ	vica	in
ਨਾਲ	nāla	with
ਹੈ	hai	is

4.1.3 Removal of Punctuation

Punctuation symbols such as ";", ",", ".", ":", "-", "'" etc. are necessarily be removed from the Punjabi sentences.

4.1.4 Word Tokenization

Every sentence is tokenized into words for the further operation such as stop-words removal and feature extraction.

4.1.5 Stop-word Elimination

Stop-words do not convey significant meaning to the sentence, and so, are removed from the Punjabi text as is seen in Table 1.

4.1.6 Word Stemming

The goal of stemming is to get the word into its basic form from its variant inflections and derivational forms as is shown in Table 2.

4.1.7 Normalization

Words need to be normalized as there are several spelling mistakes in Punjabi as is shown in Table 3.

4.2 Processing Phase

After the pre-processing phase, cleaned Punjabi text corpus is obtained over which the processing phase is applied to extract various statistical and linguistic features, for example- headline, similarity with title, length of sentence, TF-ISF, named entity recognition, cue phrase, and English-Punjabi common nouns. Each of these features is discussed in details here.

Table 2. Punjabi words and their inflected forms

Punjabi Word	Gender	Inflected Words			Number
		Inflection	Transliteration	English Translation	
ਸੋਹਣਾ sōhaṇā	Masculine	ਸੋਹਣਾ	sōhaṇā	beautiful	Singular
		ਸੋਹਣੇ	sōhaṇē		Plural
		ਸੋਹਣਿਆਂ	sōhṇiāṃ		Oblique Plural
ਸੋਹਣੀ sōhṇī	Feminine	ਸੋਹਣੀ	sōhṇī		Singular
		ਸੋਹਣੀਆਂ	sōhṇīāṃ		Plural

Table 3. Punjabi words and their spelling variations

Punjabi Words		Spelling Variations		English Translation
Word	**Transliteration**	**Variation**	**Transliteration**	
ਹਨੂਮਾਨਗੜ੍ਹ	hanūmānagaṛha	ਹਨੂਮਾਨਗੜ੍	hanūmānagaṛ	hanumangarh
ਖ਼ਿਆਲ	kẖi’āla	ਖਿਆਲ	khiāl	idea

4.2.1 Headline Feature

The headline feature (*hl*) of a text document conveys the central theme of the document. For example:
Punjabi Text: ਆਈ. ਸੀ. ਐੱਸ. ਈ 10ਵੀ ਤੇ 12ਵੀ ਦੇ ਨਤੀਜਆਿਂ ਦਾ ਐਲਾਨਅੱਜ
Transliteration: I.C.S.E. 10vīṁ tē 12vīṁ dē natīji’āṁ dā ailāna aja
English Translation: I.C.S.E results of 10th and 12th to be declared today

4.2.2 Title Similarity Feature

The title similarity feature (*ts*) represents that the word(s) within a sentence if exist in the text title also, then the sentence is quite relevant to the Punjabi text document. The score of the title similarity is computed using Equation (1):

$$ts = \frac{number\ of\ title\ words\ in\ sentence\ S}{Total\ number\ of\ words\ in\ title} \tag{1}$$

4.2.3 Sentence Length Feature

The sentence length feature (*sl*) refers to the total number of words in a Punjabi sentence. Longer sentence has more likelihood to include vital information; however, extremely shorter sentence has lesser information and are usually not incorporated within summary. Sentence length is calculated using equation (2):

$$sl = \frac{number\ of\ words\ in\ sentence\ S}{Total\ number\ of\ words\ in\ longest\ sentence} \tag{2}$$

4.2.4 Term Frequency-Inverse Sentence Frequency Feature

The TF-ISF feature (*ff*) extracts keywords from the Punjabi text using Equation (3):

$$TF\text{-}ISF\ (x) = TF\ (x) * ISF\ (x) \tag{3}$$

where:

TF(x): frequency of a word within the Punjabi sentence
ISF(x): log (N/N_i)
N: total number of sentences in the Punjabi text
N_i: number of sentences that contain the word *x*

Table 4. Named entities for the sample Punjabi text

Punjabi Text	Transliteration	English Translation	Named Entity
ਡੋਨਾਲਡ ਟਰੰਪ	Ḍōnālaḍ aṭarapa	Donald Trump	Person
ਨਰਦਿਰ ਮੋਦੀ	Naridara mōdī	Narendra Modi	Person
ਕਸ਼ਮੀਰ	Kaśamīra	Kashmir	Location
ਭਾਰਤ	Bhārata	India	Location
ਪਾਕਸਿਤਾਨ	Pākisatāna	Pakistan	Location

4.2.5 Named Entity Recognition Feature

The named entity recognition feature (*ne*) extracts the NEs- named entities from the Punjabi text, for example- person, location, organization. NEs in the Punjabi are extracted using rule-based and gazetteer lists (Gupta and Lehal, 2011a). Consider an example of the Punjabi sentence:

Punjabi Sentence: ਡੋਨਾਲਡ ਟਰੰਪ ਨੇ ਦਾਅਵਾ ਕੀਤਾ ਕਿ ਨਰਦਿਰ ਮੋਦੀ ਨੇ ਉਨ੍ਹਾਂ ਨੂੰ ਕਸ਼ਮੀਰ 'ਤੇ ਭਾਰਤ ਅਤੇ ਪਾਕਸਿਤਾਨ ਵਚਿਾਲੇ ਵਚਿੋਲਗੀ ਕਰਨ ਲਈ ਕਹਿਾ

Transliteration: Ḍōnālaḍa ṭarapa nē dā'avā kītā ki Naridara mōdī nē unhāṁ nū kaśamīra 'tē bhārata atē pākisatāna vicālē vicōlagī karana la'ī kihā

English Translation: Donald Trump claims that Narendra Modi asked him to mediate between India and Pakistan over Kashmir

In the above sentence, following Named entities are recognized as is seen in Table 4.

4.2.6 Cue Phrase Feature

The cue phrase feature (*cp*) signals semantic relations within the Punjabi text. For example, ਅੰਤਵਚਿ *atavica 'finally'*, ਸੰਖੇਪਵਚਿ *sakhēpavica 'in short'*

4.2.7 English-Punjabi Common Noun Feature

The English-Punjabi common noun feature (*ep*) represents that common English nouns are written in Punjabi as is seen in Table 5.

In the processing stage, all the features are identified and sentence scores are calculated. To do so, weights are assigned to each feature and informative score for i^{th} sentence is obtained using equation (4) and equation (5) respectively.

f_i: statistical or linguistic based Punjabi text feature

w_i: feature weight estimated via regression (Gupta and Lehal, 2011b)

n: total number of features

Table 5. English-Punjabi common nouns

English Noun	Punjabi Word	Transliteration
Technology	ਟੇਕਨਾਲੋਜੀ	Ṭēkanālōjī
Gurudwara	ਗੁਰੂਦਵਾਰਾ	Gurūdavārā
University	ਯੂਨੀਵਰਸਟਿੀ	Yūnīvarasiṭī
Gurbani	ਗੁਰਬਾਣੀ	Gurabāṇī

$$\text{InfoScore}(S_i) = \sum_{i=0}^{n} w_i * f_i \tag{4}$$

In other words,

$$InfoScore(S_i) = w_{1*}(hl) + w_{2*}(ts) + w_{3*}(sl) + w_{4*}(ff) + w_{5*}(ne) + w_{6*}(cp) + w_{7*}(ep) \tag{5}$$

4.3 PSO ALGORITHM

PSO algorithm for Punjabi text summarization contains the below mentioned steps:
Step 1- Initialization of Population
Initialization of population step finds the value of fitness for every particle within the population via fitness function.
Step 2- Current Best Solution
This step finds the current best solution that contains the best fitness value.
Step 3- Velocity Update and Position Update
For every particle, value of velocity as well as its position gets updated at this step, while using the present velocity, best position, and information from the neighborhood. The velocity and positions values get updated using equation (6) and equation (7) respectively.

$$\vec{v}_i = \vec{v}_i + \phi_1 \otimes (\vec{p}_i - \vec{x}_i) + \phi_2 \otimes (\vec{g}_g - \vec{x}_i) \tag{6}$$

$$\vec{x}_i = \vec{x}_i + \vec{v}_i \tag{7}$$

Where,

$\vec{v}_i$: velocity of ith particle
ϕ_1 : $t_1 r_1$
ϕ_2 : $t_2 r_2$
$\vec{p}_i$: personal best (best position of i^{th} particle found so far)
$\vec{x}_i$: current position of i^{th} particle in the swarm
$\vec{g}_g$: global best (best position from particle neighborhood)
$\otimes$: vector multiplication
t_1: personal acceleration coefficient
r_1: first vector of random numbers uniformly chosen from [0, 1]
t_2: social acceleration coefficient
r_2: second vector of random numbers uniformly chosen from [0, 1]

Step 4- Stopping Criteria

Table 6. PSO parameters setting over Punjabi_Dataset_I and Punjabi_Dataset_II

Parameter	Value	Dataset	Value	Dataset
Number of maximum iterations	50	Punjabi_Dataset_I	20	Punjabi_Dataset_II
Size of Population	20		10	
Personal Acceleration Coefficient	2		2	
Social Acceleration Coefficient	2		2	

Check for the stopping criteria. If the criterion is met then particles are ranked and return the best result, otherwise process is repeated again from Step 2.

For the fitness function, computation on the informative scores is performed for those sentences that are part of the generated summary using Equation (8).

$$F_{\alpha} = \sum_{l=1}^{k} InfoScore(S_l) \tag{8}$$

Where,

F_{α} : fitness value of α^{th} candidate summary

k: size of candidate summary

The crux for PSO algorithm for Punjabi text summarization task is that PSO selects the maximum informative sentences where every particle updates its velocity and position as per its current velocity, best position and information of its neighbors'. As the number of iterations progresses, population turns out to be homogeneous, and so, the best score does not fluctuate for a number of steps. At this point, Punjabi text summary with the best fitness function value is effectively produced.

5. EXPERIMENTAL SETUP, DATASETS AND RESULTS

5.1 Experimental Setup

To work upon the proposed methodology, installation of python[@4] version 3.5 or higher is required to ensure that it is compatible with the Punjabi language text. Unlike the English language which supports American Standard Code for Information Interchange (ASCII) format, the Punjabi language deals with the Unicode format. Thus, the Punjabi text document is to be saved with the UTF-8[5] encoding. In addition, several python libraries are needed, for example, NumPy[6]: numeric python, pandas[7]: data analysis, lxml[8]: web scrapping, and psopy[9]: SciPy[10] compatible PSO implementation.

5.2 Punjabi Datasets

The proposed summarization system for the Punjabi text is evaluated over two datasets. First dataset, we have named as Punjabi_Dataset_I is downloaded as monolingual Punjabi corpus from Indian Languages Corpora Initiative Phase-II (ILCI Phase-II)[11] under the initiative of Ministry of Electronics and Information Technology (MeitY), Government of India. The Punjabi corpus comprises of approximately 30,000 sentences of general domain having features of unique ID, UTF-encoding and text file format.

Second dataset, we have named as Punjabi_Dataset_II is downloaded as Punjabi-Hindi[12] parallel corpus that is produced by human translators. We have considered the Punjabi text out of the entire corpus which comprises of a total of 1,000 sentences from the Web having tourism/travel domain and text encoding is UTF-8.

Table 7. Comparison of ROUGE measures over Punjabi_Dataset_I and Punjabi_Dataset_II

Dataset	ROUGE	Precision	Recall	F-Measure
Punjabi_Dataset_I	ROUGE-1	0.6904	0.6731	0.6816
	ROUGE-2	0.5892	0.5442	0.5658
Punjabi_Dataset_II	ROUGE-1	0.7836	0.7957	0.7896
	ROUGE-2	0.75823	0.7754	0.7667

5.3 Evaluation Criteria

ROUGE (Gupta and Kaur, 2016) measure i.e. Recall-Oriented Understudy for Gisting Evaluation which is developed by Document Understanding Conferences- NIST is considered to measure the summary quality. Both ROUGE-1 and ROUGE-2 tests are used over each of the datasets- Punjabi_Dataset_I and Punjabi_Dataset_II respectively. Table 6 represents setting of PSO parameters for these datasets one by one.

Table 7 gives comparison of ROUGE measures (ROUGE-1, ROUGE-2) over both the Punjabi datasets. ROUGE-1 has higher value as compared to ROUGE-2, for both of the datasets (Punjabi_Dataset_I and Punjabi_Dataset_II) since ROUGE-2 finds the word pairs, word sequences etc. within the datasets; however, Rouge-1 is independent of such word pairs. Also, ROUGE measures of Punjabi_Dataset_II is comparatively higher (especially higher recall) than that of the Punjabi_Dataset_I since Punjabi_Dataset_II is domain specific with quite lesser number of sentences; however, Punjabi_Dataset_I is of general domain with much larger number of sentences. The highest measure, ROUGE-1 is achieved for the parallel corpus with precision, recall and F-measure as 0.7836, 0.7957 and 0.7896 respectively.

6. CONCLUSION

Particle Swarm Optimization (PSO) algorithm for Text Summarization (TS) task is carried out for the Punjabi language using several python libraries- NumPy, pandas, lxml, psopy, SciPy. PSO selects the maximum informative sentences from Punjabi corpus where for every particle there is an update of its velocity as well as an update of its position; as per its current velocity, the best position and information of its neighbors. As the number of iterations progresses, population turns out to be homogeneous and the best score does not fluctuate further. And then, summary of the Punjabi text with the best fitness function value is effectively produced. To do so, pre-processing and processing phases of TS are incorporated. The pre-processing phase cleans the Punjabi text document while entailing features such as input restriction, sentence tokenization, removal of punctuation, word tokenization, stop-word elimination, word stemming, and normalization. On the other hand, the processing phase is applied to extract various statistical and linguistic features such as headline, title similarity, sentence length, TF-ISF, named entity, cue phrase, and English-Punjabi common noun. PSO algorithm over the Punjabi text-based TS task is tested over two different datasets- Punjabi_Dataset_I and Punjabi_Dataset_II. Both these datasets are of different genres and are evaluated over ROUGE measures. ROUGE-1 gives better results in comparison to ROUGE-2 over both the datasets. And, the best result is for ROUGE-1 over Punjabi_Dataset_II having precision as 0.7836, recall as 0.7957 and F-measure as 0.7896 respectively.

In future, more complex features can be added to the processing phase for deeper understanding of the text and to generate abstractive summary. Also, work can be extended to make the TS methodology as language and platform independent too.

REFERENCES

Abuobieda, A., Salim, N., & Eltayeb, R. A., bin Wahlan, M. S., Suanmali, L., & Hamza, A. (2011). Pseudo Genetic and Probablisitic-based Feature Selection Method for Extractive Single Document Summarization. *Journal of Theoretical and Applied Information Technology*, *32*(1), 80–87.

Ahmad, M. T., Malik, M. K., Shahzad, K., Aslam, F., Iqbal, A., Nawaz, Z., & Bukhari, F. (2020). Named Entity Recognition and Classification for Punjabi Shahmukhi. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *19*(4), 1–13.

Al-Abdallah, R. Z., & Al-Taani, A. T. (2017). Arabic Single-Document Text Summarization Using Particle Swarm Optimization Algorithm. *Procedia Computer Science*, *117*, 30–37.

Al-Zahrani, A. M., Mathkour, H., & Abdalla, H. I. (2015). PSO-Based Feature Selection for Arabic Text Summarization. *Journal of Universal Computer Science*, *21*(11), 1454–1469.

Almazaydeh, L. (2018). Automatic Arabic Text Summarization System (AATSS) Based on Morphological Analysis. *International Journal of Intelligent Systems Technologies and Applications*, *17*(3), 272–280.

Binwahlan, M. S., Salim, N., & Suanmali, L. (2009). Swarm Based Features Selection for Text Summarization. *International Journal of Computer Science and Network Security*, *9*(1), 175–179.

Gambhir, M., & Gupta, V. (2017). Recent Automatic Text Summarization Techniques: A Survey. *Artificial Intelligence Review*, *47*(1), 1–66.

Gunawan, D., Pasaribu, A., Rahmat, R. F., & Budiarto, R. (2017, April). Automatic Text Summarization for Indonesian Language Using Textteaser. *IOP Conference Series. Materials Science and Engineering*, *190*(1), 1–6.

Gupta, J. P., Tayal, D. K., & Gupta, A. (2011). A TENGRAM Method Based Part-of-Speech Tagging of Multi-Category Words in Hindi Language. *Expert Systems with Applications*, *38*(12), 15084–15093.

Gupta, V., & Kaur, N. (2016). A Novel Hybrid Text Summarization System for Punjabi Text. *Cognitive Computation*, *8*(2), 261–277.

Gupta, V., & Lehal, G. S. (2011a). Named Entity Recognition for Punjabi Language Text Summarization. *International Journal of Computers and Applications*, *33*(3), 28–32.

Gupta, V., and Lehal, G. S. (2011b). Automatic Keywords Extraction for Punjabi Language. *International Journal of Computer Science Issues, 8*(5), 327-330.

Jain, A. (2019). *Named Entity Recognition for Hindi Language Using NLP Techniques (PhD Thesis)*. Jaypee Institute of Information Technology. https://shodhganga.inflibnet.ac.in/handle/10603/241558

Jain, A., & Arora, A. (2018a). Named Entity System for Tweets in Hindi Language. *International Journal of Intelligent Information Technologies*, *14*(4), 55–76.

Jain, A., Gairola, R., Jain, S., & Arora, A. (2018a). Thwarting Spam on Facebook: Identifying Spam Posts Using Machine Learning Techniques. In Social Network Analytics for Contemporary Business Organizations, (pp. 51-70). IGI Global.

Jain, A., Gupta, A., Sharma, N., Joshi, S., & Yadav, D. (2018b, April). Mining Application on Analyzing Users' Interests from Twitter. *Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT)*, 26-27.

Jain, A., Tayal, D. K., & Arora, A. (2018c). OntoHindi NER – An Ontology Based Novel Approach for Hindi Named Entity Recognition. *International Journal of Artificial Intelligence*, *16*(2), 106–135.

Jain, A., Tayal, D. K., Yadav, D., & Arora, A. (2020). Research Trends for Named Entity Recognition in Hindi Language. In *Data Visualization and Knowledge Engineering* (pp. 223–248). SPRINGER.

Jain, A., Vishnoi, P., Kumar, H., & Saad, S. (2013, October). Web Advertisement Image Filtration Using Internet Image Advertisement Blocker Tool. *Proceedings of the International Conference on Electronics, Communication and Information Technology*, 1-3.

Jain, A., Yadav, D., & Tayal, D. K. (2014, September). NER for Hindi Language Using Association Rules. In *Proceedings of the International Conference on Data Mining and Intelligent Computing (ICDMIC)*, (pp. 1-5). IEEE.

Kaur, A., & Gupta, V. (2014). Proposed Algorithm of Sentiment Analysis for Punjabi Text. *Journal of Emerging Technologies in Web Intelligence*, *6*(2), 180–183.

Kaur, K., & Gupta, V. (2011). Topic Tracking for Punjabi Language. *Computer Science Engineering International Journal*, *1*(3), 37–49.

Kaur, M., & Kaur, J. (2013). Deadwood Detection and Elimination in Text Summarization for Punjabi Language. *International Journal of Engineering Science*, *8*, 51–59.

Kaur, R., Sharma, R. K., Preet, S., & Bhatia, P. (2010). Punjabi WordNet Relations and Categorization of Synsets. *3rd National Workshop on IndoWordNet Under the Aegis of the 8th International Conference on Natural Language Processing (ICON 2010)*.

Lins, R. D., de Mello, R. F., & Simske, S. J. (2020). DocEng'2020 Competition on Extractive Text Summarization. *Proceedings of the ACM Symposium on Document Engineering 2020*, 1-4.

Meena, Y. K., & Gopalani, D. (2016). Statistical Features for Extractive Automatic Text Summarization. In Enterprise Big Data Engineering, Analytics, and Management, (pp. 126-144). IGI Global.

Narang, A., Sharma, R. K., & Kumar, P. (2013). Development of Punjabi WordNet. *CSI Transactions on ICT, 1*(4), 349-354.

Nidhi, V. G. (2012, December). Domain Based Classification of Punjabi Text Documents. *Proceedings of COLING*, 297-304.

Saggion, H., & Poibeau, T. (2013). Automatic Text Summarization: Past, Present and Future. In Multi-Source, Multilingual Information Extraction and Summarization, (pp. 3-21). Springer.

Sarkar, S., Saha, S., Bentham, J., Pakray, P., Das, D., & Gelbukh, A. F. (2016). NLP-NITMZ@ DPIL-FIRE2016: Language Independent Paraphrases Detection. FIRE (Working Notes), 256-259.

Sharma, D., & Jhajj, P. (2010). Recognition of Isolated Handwritten Characters in Gurmukhi Script. *International Journal of Computers and Applications*, *4*(8), 9–17.

Sharma, S. K. (2017). Clauses Detection in Punjabi Language. *International Journal of Innovations & Advancement in Computer Science, 6*(8).

Sharma, S. K. (2019). Sentence Reduction for Syntactic Analysis of Compound Sentences in Punjabi Language. *EAI Endorsed Transactions on Scalable Information Systems*, *6*(20), e4.

Singh, A., & Singh, P. (2015). Punjabi Dialects Conversion System for Malwai and Doabi Dialects. *Indian Journal of Science and Technology*, *8*(27), 1–6.

Singh, J., Singh, G., Singh, R., & Singh, P. (2018). Morphological Evaluation and Sentiment Analysis of Punjabi Text Using Deep Learning Classification. *Journal of King Saud University-Computer and Information Sciences*. 10.1016/j.jksuci.2018.04.003

Singh, P., & Kaur, R. (2014). A Review on Question Generation System from Punjabi Text Contain Historical Information. *International Journal of Computer Science and Mobile Computing*, 185-189.

Song, S., Huang, H., & Ruan, T. (2019). Abstractive Text Summarization Using LSTM-CNN Based Deep Learning. *Multimedia Tools and Applications*, *78*(1), 857–875.

Verma, A., Morato, J., Jain, A., & Arora, A. (2020). Relevant Subsection Retrieval for Law Domain Question Answer System. In *Data Visualization and Knowledge Engineering* (pp. 299–319). Springer.

Zhang, J., Zhang, X., & Yang, J. (2020). Multiobjective Particle Swarm Community Discovery Arithmetic Based on Representation Learning. *Concurrency and Computation*, e5788.

Zhang, X., Zhang, R., Wang, J., & Wang, L. (2018). An Adaptive Particle Swarm Optimization Algorithm Based on Aggregation Degree. *Recent Advances in Electrical & Electronic Engineering (Formerly Recent Patents on Electrical & Electronic Engineering)*, *11*(4), 443-448.

ENDNOTES

1 http://punjabipos.learnpunjabi.org/
2 http://www.cfilt.iitb.ac.in/indowordnet/
3 http://anoopkunchukuttan.github.io/indic_nlp_library/
4 https://www.python.org/
5 https://www.gurbani.org/unicode.php
6 https://numpy.org/
7 https://pandas.pydata.org/
8 https://lxml.de/
9 https://pypi.org/project/psopy/
10 https://www.scipy.org/
11 https://tdil-dc.in/index.php?option=com_download&task=showresourceDetails&toolid=1890&lang=en
12 https://tdil-dc.in/index.php?option=com_download&task=showresourceDetails&toolid=1264&lang=en

Arti Jain is working as Assistant Professor (Sr. Grade) in the Department of Computer Science & Engineering at Jaypee Institute of Information Technology (JIIT), Noida (UP), India. She is having academic experience of 18 years.She is PhD (CSE) from JIIT Noida. She is member of IEEE, INSTICC, IAENG, IASSE, IFERP, and Life Member of TERA. She has more than 20 research papers in peer-reviewed International Journals, Book Chapters, and International Conferences. She has supervised one M.Tech Thesis and around B.Tech projects. Currently she is supervising 1 Ph.D. candidate in the area of Social Network Analysis. She is reviewer of reputed and peer-reviewed International Journals- Taylor & Francis, IGI Global, Wiley, TISA and Inderscience. She is TPC member and reviewer of several International Conferences- SCES, ICDSAA, UPCON 2020, CONFLUENNCE 2020, BigDML, ComITCon 2019, UPCON 2018-19. She is editorial board member of American Journal of Neural Networks and Applications. She is special session organizer in International Conference on Innovative Computing and Communication (ICICC 2020), New Delhi, India. She has participated in more than 70 specialized short-term courses. Her research interest includes Natural Language Processing, Machine Learning, Data Science, Deep Learning, Social Media Analysis and Data Mining.

Divakar Yadav is working as Associate Professor in the Department of Computer Science & Engineering at National Institute of Technology (NIT), Hamirpur (HP), India. He did his undergraduate in Computer Science & Engineering (1999), Post Graduate in Information Technology (2005) and PhD in Computer Science & 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.

Anuja Arora is working as Associate Professor in the Department of Computer Science & Engineering at Jaypee Institute of Information Technology (JIIT), Noida (UP), 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, etc. She is also the reviewer of various Springer, IGI Global, Inderscience, and De Gruyter Journals. She has guided more than 17 M.Tech Thesis and around 100 B.Tech Projects.