Textual Entailment based Figure Summarization for Biomedical Articles

NAVEEN SAINI*, Indian Institute of Technology Patna, India
SRIPARNA SAHA, Indian Institute of Technology Patna, India
PUSHPAK BHATTACHARYYA, Indian Institute of Technology Patna, India
HIMANSHU TUTEJA, Birla Institute of Technology Mersa, India

The current paper proposes a novel unsupervised approach (FigSum++) for automatic figure summarization in biomedical scientific articles using a multi-objective evolutionary algorithm. The problem is treated as an optimization problem where relevant sentences in the summary for a given figure are selected based on various sentence scoring features (or objective functions): the textual entailment score between sentences in the summary and figure's caption, the number of sentences referring to that figure, semantic similarity between sentences and figure's caption, the number of overlapping words between sentences and figure's caption etc. These objective functions are optimized simultaneously using multi-objective binary differential evolution (MBDE). MBDE consists of a set of solutions and each solution represents a subset of sentences to be selected in the summary. MBDE generally uses single DE variant, but, in the current study, ensemble of two different DE variants measuring diversity among solutions and convergence towards global optimal solution, respectively, is employed for efficient search. Usually, in any summarization system, diversity amongst sentences (called as anti-redundancy) in the summary is a very critical feature and it is calculated in terms of similarity (like cosine similarity) among sentences. In this paper, a new way of measuring diversity in terms of textual entailment is proposed. To represent the sentences of the article in the form of numeric vectors, recently proposed, BioBERT, a pre-trained language model in biomedical text mining is utilized. An ablation study has also been presented to determine the importance of different objective functions. For evaluation of the proposed technique, two benchmark biomedical datasets containing 91 and 84 figures, respectively, are considered. Our proposed system obtains 5% and 11% improvements in terms of F-measure metric over two datasets, respectively, in comparison to the state-of-the-art unsupervised methods.

CCS Concepts: • **Information systems** → **Information extraction**; **Summarization**.

Additional Key Words and Phrases: Figure-assisted text summarization, textual entailment, evolutionary computing, multi-objective optimization (MOO).

ACM Reference Format:
Naveen Saini, Sriparna Saha, Pushpak Bhattacharyya, and Himanshu Tuteja. 2019. Textual Entailment based Figure Summarization for Biomedical Articles . 1, 1, Article 1 (January 2019), 23 pages. https://doi.org/10.1145/3357334

1 INTRODUCTION

Automatic summarization [18] focuses on shortening a given text/image maintaining the crux of the information as in the given input. The ability to simplify the information has brought attention to this area. Summarization can assist

*Corresponding author

Authors' addresses: Naveen Saini, Indian Institute of Technology Patna, Bihar, India, naveen.pcs16@iitp.ac.in; Sriparna Saha, Indian Institute of Technology Patna, Bihar, India, sriparna@iitp.ac.in; Pushpak Bhattacharyya, Indian Institute of Technology Patna, Bihar, India, pb@iitp.ac.in; Himanshu Tuteja, Birla Institute of Technology Mersa, Jaipur, India, tuteja.himanshututs@gmail.com.

Manuscript submitted to ACM

many application areas such as search results, shortening the medical reports, news articles etc. Due to the rapid increase in the text data, the task to summarize them into the shorter form was in huge demand [20, 34]. Over the last decade, automatic summarization is one of the principal, challenging issues in Natural Language Information Processing [21]. There exist extractive summarization systems for different tasks such as microblog summarization [10], single [34] and multi-document [3] summarization, etc. Summarization can be classified into two types: extractive and abstractive. Extractive [34] generates a summary by selecting the sentences from the document. But, abstractive [29] has the freedom to explore the words/sentences which aren't present in the text document. It requires reconstruction of the sentences.

In this paper, we introduce a novel extractive summarization technique to deal with the problem of summarizing the figures in biomedical articles in an unsupervised way. According to Futrelle[13], 50% of the texts in biomedical articles are related to figures only. Moreover, as per [45], only caption of the figure and title of the article with abstract convey 30% of the information related to the figure. These figures are always difficult to interpret by humans as well as machines. Therefore, associated texts in the article can be used to describe them. For example- Agarwal and Yu [2] proposed a system, *FigSum*, to generate summary of images related to biomedical domain using the scattered text throughout the various sections of the scientific articles like the introduction, proposed method, results and so on. The top scoring sentences having high tf-idf cosine similarity [26] with the figure's caption and article's main theme were considered as a part of the summary. But, in a biomedical article, a number of sentences are there and it is difficult to decide which are more relevant to the figure. Therefore, there is a need to develop a more sophisticated system which summarizes figures by extracting the relevant sentences by optimizing different criteria in an unsupervised way.

To measure the similarity between sentences, a well known measure, cosine similarity [16] is used. Higher the similarity, more close they are. But, it requires vector representation of the sentences for which recently developed, a pre-trained language model on a large biomedical corpora, namely, BioBERT [17], is utilized. Note that it is capable of capturing the semantic similarity between sentences.

1.1 Motivation

Our work is motivated by the fact that in a biomedical article, many sentences are there, and those may be relevant to the figure with respect to different perspectives (also called scoring features or fitness functions or objective functions) like whether the sentences refer to that figure (SRF), amount of similarity the sentences have with figure's caption (SFC), number of 1-gram overlapping words between sentence and figure's caption (SOC1), number of 2-gram overlapping words between sentence and figure's caption (SOC2). Moreover, whether a sentence entails to figure's caption (STE) or not, can be considered as another scoring function. Therefore, in our proposed system (FigSum++), these sentence scoring functions are optimized simultaneously in an unsupervised way using the multi-objective (MOO) binary differential evolution [42] algorithm (MBDE) which is an evolutionary algorithm (EA). However, some other optimization strategies like AMOSA [5], PSO [46], etc. also exist in the literature. But, DE is preferred because of its better performance compared to others [31–34]. To avoid redundancy in summary between sentences, another goodness measure named as anti-redundancy (SAR) is also taken into account in our optimization process. Note that SAR employs the cosine similarity while computing the similarity/dissimilarity between sentences of the summary in semantic space. It is also important to note that SAR is considered to maintain diversity among-st sentences.

MBDE [41] is a population-based meta-heuristic optimization algorithm which starts it's search with a set of solutions (or, chromosomes, used interchangeably) called as population. Each solution is represented as the binary vector denoting

a set of possible sentences to be selected in the generated summary. Generally, in MBDE framework, rand/1/bin scheme/variant is used to generate a new solution at each iteration using fix values of two parameters, mutant factor (F) and crossover rate (CR) [39]. As a result, the search ability of these algorithms could be limited. Note that in the MBDE framework, CR and F, are the two crucial parameters which help in reaching the global optimal solution. Moreover, rand/1/bin scheme may not be efficient as it has exploratory nature. But, the best solution (or the best summary for a given figure) may lie in local or global region. Therefore, in this paper, instead of rand/1/bin, the ensemble of two other DE schemes namely, current-to-rand/1/bin and current-to-best/1/bin, is used in the new solution generation process. Motivation behind using these variants is that in any evolutionary algorithm, diversity among solutions and convergence towards true/global optimal solutions are the important phenomena which can be achieved using current-to-rand/1/bin and current-to-best/1/bin, respectively. More information about these variants can be found in the paper [39]. Also, to get rid of fixing the values of F and CR parameters, a pool of values of these parameters are also considered based on literature survey [39, 42]. These DE variants can randomly select F and CR values from the given pool. This phenomenon is shown in Figure 3 (more description provided in section 4.4).

As it is difficult to decide which set of objective functions is the best suited for our task using MOO algorithm, an ablation study has also been done on the selected objective functions. Here, ablation study means various combinations of the objectives functions, for example, (a) SAR_TE and STE; (b) SAR_TE and SRF; (c) SAR_TE, SRF, and SFC, etc., are optimized simultaneously using MBDE framework in different runs of our proposed algorithm.

Textual entailment (TE) [28] is itself a challenging problem in NLP domain. The importance of TE can be understood by the BioNLP[1] 2019 shared task on textual inference and question entailment on biomedical text. Definition of TE states that a sentence 'p' (called as hypothesis) is said to be entailed by sentence 'q' (called as premise) if 'p' can be inferred from 'q' [28]. It also describes whether relationship between 'p' and 'q' is contradictory or neutral. An example of entailment taken from MedNLI[2] dataset is shown below:

p : *Patient had aphasia.*
q : *Patient was not able to speak, but appeared to comprehend well.*

where, 'p' is entailed by 'q' and represented as $q \rightarrow p$. Due to the popularity of TE, in this paper, we have proposed a different way of measuring anti-redundancy in summary. The sentences, in summary, should not be entailed to each other to maintain diversity among-st sentences. It's mathematical formulation is described in Section 2. Thus, in total, two ways of measuring anti-redundancy in summary are explored: one makes use of cosine similarity, while, another makes use of textual entailment relationship between sentences.

1.2 Contributions

Following are the major contributions of this paper:

(1) To the best of our knowledge, the proposed work is the first attempt in developing a multi-objective based framework for solving figure-summarization task in which various sentence scoring features like the number of sentences referring to the figure, semantic similarity between sentences and figure's caption, the number of overlapping words between sentences and figure's caption etc. are optimized simultaneously to generate a good

[1]https://aclweb.org/aclwiki/BioNLP_Workshop
[2]https://physionet.org/physiotools/mimic-code/mednli/

quality summary. Moreover, whether, sentences in summary, entail to figure's caption or not, also considered as another objective function in the optimization process.

(2) Any multiobjective algorithm should satisfy two properties: diversity among solutions and convergence towards true Pareto optimal front. To achieve the same, two different DE variants (current-to-rand/1/bin and current-to-best/1/bin) are utilized in the current framework. The first schema incorporates diversity and the second one includes convergence.

(3) To minimize redundancy among-st sentences in the generated summary, a new method utilizing textual entailment relationships between sentences is proposed.

(4) To measure the similarity among sentences in the semantic space, a deep learning-based recently proposed pre-trained language model, namely, BioBERT [17] developed for biomedical text mining, is utilized.

(5) To find out the set of most contributing objective functions in our optimization process, ablation study is presented.

(6) All the existing approaches provide a single fixed length summary (depending on the user). But, as our approach is based on population-based strategy, therefore, multiple summaries of different lengths are provided to the end-user and the user can select any summary based on his/her choice.

We have tested our system on two gold-standard datasets, *FigSumGS1* and *FigSumGS2* containing 91 and 84 figures, respectively. Results obtained clearly show the superiority of our proposed algorithm in comparison to various state-of-the-art techniques. The organization of the paper is as follows: Section 2 discusses the literature survey and background knowledge. Section 3 and 4 discusses the problem definition and methodology of the proposed architectures used for figure-summarization, respectively. Experimental setup is presented in Sections 4 followed by results discussion in Sections 6. Finally, the paper is concluded in Section 7.

2 RELATED WORKS AND BACKGROUND KNOWLEDGE

In the literature, a large number of works exist on summarization of text documents/scientific articles [20, 34]. We have found mainly four categories of works done on text document summarization till now: (a) supervised; (b) unsupervised; (c) neural-network-based; and, (d) meta-heuristic. Supervised techniques such as SVM [44] etc., make use of labeled data for training (i.e., whether sentence belongs to the summary or not) which requires manual effort and is a time-consuming step. Some other papers are [23, 38]. On the other hand, unsupervised methods don't require labeled data. Some of the works using unsupervised methods are [9, 11]. The methods based on meta-heuristic strategies, developed in the papers [4, 36], utilized different types of optimization techniques like PSO (Particle Swarm Optimization) [46], NSGA-II (non-dominated sorting genetic algorithm) [8] etc. to optimize the summary quality. Some deep learning based models like RNN (recurrent neural network) [20] are also developed in the literature for document summarization task. But, a few works have been reported on summary generation of figures from the text documents/articles. The authors of the papers [1] [2] [12] [14] [25] [43] have carried out works in the same domain. The contributions of these works are provided in the Table 1. To the best of our knowledge, there is no other work after that work. Moreover, the existing technique doesn't make use of multi-objective optimization approach to solve the figure-summarization task.

2.1 BioBERT Language Representation

BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) is a domain specific language representation. It was trained on large-scale biomedical corpora. The model was applied on different NLP

Table 1. Descriptions of different existing methods for Figure-summarization

Existing paper	Contribution
Passonneau et al. [24]	Proposed a system to summarize the workflow diagrams. The major drawback was that it requires a list of attribute values describing the diagrams.
Futrelle [12]	Proposed the idea of figure summarization and discussed various challenges and issues related to it.
Futrelle [13]	Authors used structure of the diagram, the text of the figure's caption and text in the article for summarizing figures.
Afantenos et al. [1]	Discussed about various summarization techniques that can be used in bio-medical articles.
Agarwal et al. [2]	Proposed a system, *FigSum*, to summarize images of biomedical articles; authors assume that figure's information was scattered throughout the article; the sentences with high tf-idf [35] cosine similarity [26] with the figure's caption and article's main theme were considered as a part of the summary.
Peng et al. [43]	Proposed the idea of summarization of information graphics and used the paragraphs in a multi-modal document related to news domain.
Bhatia et al. [6]	Authors used a supervised approach to generate figure summary by identifying the relevant sentences on the basis of similarity of sentences in the article with the figure's caption and sentences referring to that figure.
Ramesh et al. [25]	Proposed a system, *FigSum+*, an extended version of *FigSum* [2]. Authors of this paper have explored various approaches to generate the summary of bio-medical images in the scientific article. Some of the approaches are developed using surface-cue words, for example, identifying paragraphs and sentences referring to the figure.

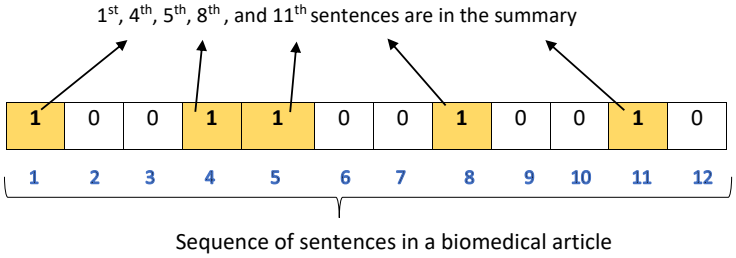


Fig. 1. ith solution representation in the population. Here, 12 is the number of sentences in the article, '0' denotes that the sentence will not be a part of extractive summary and vice-versa.

tasks and improved performance has been reported for solving many BioNLP tasks [17] such as biomedical relation extraction, biomedical named entity recognition, etc. Therefore, in our task, we have made use of this representation to represent the sentences in semantic space.

3 PROBLEM DEFINITION

Consider a biomedical article $\mathcal{A}$ consisting of N sentences, $\mathcal{A}$=$\{s_1, s_2, \ldots, s_N\}$ and a set of M figures {Fig-1, Fig-2, …, Fig-M}. We aim to summarize mth figure (Fig-m) using these sentences. Then, our main objective is to select a subset of sentences, $S \in \mathcal{A}$, related to mth figure, defined as follows:

$$S_{min} \leq \sum_{i=1}^{N} B_i \leq S_{max} \quad and \quad B_i = \begin{cases} 1, & \text{if } s_i \in S \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Table 2. Description of symbols used in describing objective functions (mathematical formulation).

Symbol	Description
x_i	ith solution is denoted as x_i (as our system generates a set of solutions and each solution corresponds to a subset of sentences forming summary for mth figure)
N	maximum length of the solution or number of sentences in the article
x_{ij}	denotes the jth component (1/0) of ith solution; the value 0 indicates that jth sentence is not selected for summarization and 1 indicates that the sentence is selected for summarization.
s_{ij}	jth sentence of the kth article, belonging to ith solution
$\|\ .\ \|$	measures the count
$\mathcal{M}$	cosine similarity between two sentences
C_{km}	caption of mth figure in kth article
$S1$	the set of sentences in the article entailed to C_{km}
$S2$	the set of sentences in the ith solution entailed to C_{km}
$s_{ia} \rightarrow s_{ib}$	bth sentence of the kth article, belonging to ith solution entailed by ath sentence belonging to same article and same solution
$\uparrow$ and $\downarrow$	indicate fitness functions are of maximization and minimization type, respectively.

such that {SAR_TE(S), SAR_CS(S), STE(S), SFC(S), SRF(S), SOC1(S), SOC2(S)} are optimized simultaneously; where, S_{min} and S_{max} are the minimum and the maximum number of sentences to be present in the summary, respectively; SAR_TE(S), SAR_CS(S), STE(S), SFC(S), SRF(S), SOC1(S), and, SOC2(S) are the objective functions measuring different aspects/qualities of summary at syntactic and semantic level and discussed in the subsequent section. Note that (a) there can also be two or more than two objective functions instead of seven; (b) In STE, SFC, SOC1, and SOC2, mth figure's caption is utilized. Let us assume that we want to generate summary of mth figure in kth article whose caption is C_{km}. Then the steps of computing objective functions for ith solution are enumerated below and the notations used while calculating these objectives are provided in Table 2. Representation of ith solution is shown in Figure-1.

(1) SAR : There can be lot of redundant sentences in the article. Therefore, to reduce the redundancy in the summary, two versions of SAR are considered:

(a) SAR_CS ($\downarrow$): It measures the cosine similarity (CS) between sentences in the summary. Let us call it as SAR_CS. It's score for ith solution is calculated as

$$SAR\_CS = \frac{(\sum_{a,b=1,a\neq b}^{N} \mathcal{M}(s_{ia}, s_{ib}))}{O} \quad \text{if} \quad x_{ia} = x_{ib} = 1 \tag{2}$$

where, O is the total number of paired sentences considered during calculation and rest of the notations are discussed in Table 2.

(b) SAR_TE ($\downarrow$): Second version measures the anti-redundancy between sentences of the summary in terms of textual entailment relationships. It can be defined as below

$$SAR\_TE = \frac{\sum_{a=1}^{N}\sum_{b=1}^{N} Q(s_{ia}, s_{ib})}{O} \quad \text{if} \quad x_{ia} = x_{ib} = 1 \quad and \quad Q(s_{ia}, s_{ib}) = \begin{cases} 1 & \text{if } s_{ia} \rightarrow s_{ib} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Here O is the total number of paired sentences considered during calculation.

(2) STE ($\uparrow$): This function calculates the entailment relationships between sentences of the summary and figure's caption. To calculate the score for this function, first, we need to identify the sentences in the articles which are

entailed to mth figure caption, i.e., C_{km}. Let us denote this set as S. Then, the number of overlapping sentences belonging to ith solution and S is calculated which will be considered as STE score. Mathematically, it can be expressed as

$$STE = \mid S1 \cap S2 \mid \tag{4}$$

Note that to identify the entailed sentences in the article to C_{km}, we have used the pre-trained model available at https://github.com/jgc128/mednli_baseline. In this model GloVe[3] word2vec embeddings (840 B tokens, 2.2M vocabulary size, and 300-dimensional vectors) are used for initialization followed by fine tuning using fastText[4] word embedding on BioASQ[5] and MIMIC-III[6] data. Note that BioASQ is the collection of 12, 834, 585 abstracts of scientific articles related to the biomedical domain and MIMIC-III data consists of 2, 078, 705 clinical notes with 320 tokens.

(3) SFC (↑): In this objective, average cosine similarity between sentences in the ith solution and figure's caption (C_{km}) belonging to kth article is calculated. Mathematically it's score is calculated as:

$$SFC = \frac{\sum_{j=1}^{N} \mathcal{M}(s_{ij}, C_{km})}{L} \quad if \quad x_{ij} = 1 \tag{5}$$

where, L is the count of x_{ij} having value of 1.

(4) SRF (↑): It counts the number of sentences present in the ith solution referring to the mth figure by using keyword 'Figure-m'. It is computed as

$$\sum_{j=1}^{N} I_j \quad \text{where} \quad I_j = 1, \text{if sentence } s_{ij} \text{ refers to mth figure and } x_{ij} = 1 \tag{6}$$

(5) SOC1 (↑): It counts the number of 1-gram overlapping words between sentences present in the ith solution and mth figure's caption; it is defined as follows:

$$\sum_{j=1}^{N} \mid (Words \in s_{ij} \cap (Words \in C_{km}) \mid \quad \text{if} \quad x_{ij} = 1 \quad \text{and Words are in the form of 1-gram unit} \tag{7}$$

(6) SOC2 (↑): It is similar to SOC1. Only difference is that in place of 1-gram, number of 2-gram overlapping words are counted. It is calculated as below:

$$\sum_{j=1}^{N} \mid (Words \in s_{ij}) \cap (Words \in C_{km}) \mid \quad \text{if} \quad x_{ij} = 1 \quad \text{and Words are in the form of 2-gram unit} \tag{8}$$

4 PROPOSED APPROACH

This section discusses the various steps followed in our proposed approach (FigSum++). The corresponding flowchart is also shown in Figure 2. Due to length restrictions, we have provided the pseudo code of our proposed approach in the supplementary sheet.

[3]https://nlp.stanford.edu/projects/glove/
[4]https://fasttext.cc/docs/en/english-vectors.html
[5]http://participants-area.bioasq.org/general_information/Task6a/
[6]https://mimic.physionet.org/

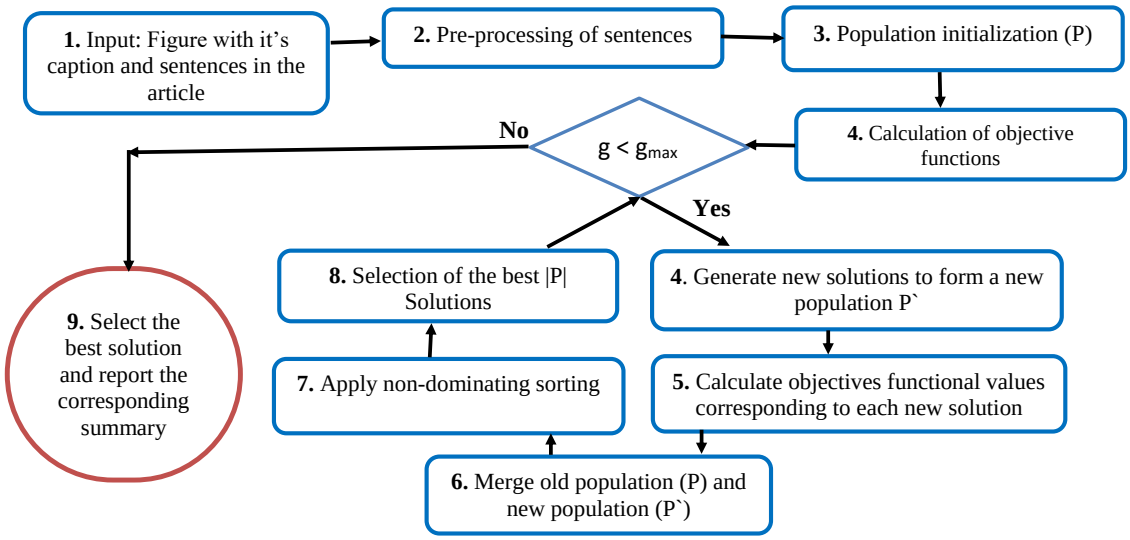


Fig. 2. Flow chart of the proposed architecture where, g is the current generation number initialized with 0 value, g_{max} is the user-defined maximum number of generations, $|P|$ is the size of the population.

4.1 Pre-processing

Before applying our proposed approach, pre-processing of biomedical article is required. List of steps followed to perform the same are described below:

(1) Biomedical articles was available in the pdf format, therefore, first, sentences are extracted using Grobid tool[7]. Note that while extracting the sentences, abstract and appendix (if available) are excluded. Only remaining sections like introduction, methodology etc. are used.
(2) Removal of stop-words.

Moreover, the cosine similarity between sentences is pre-computed as it will be required while running the experiments. To calculate the same, first, sentences are represented in the form of fixed length numeric vectors using the BioBERT [17] language model.

4.2 Population Initialization and Solution Representation

This step includes initialization of population. Population P consists of a set of solutions $< \vec{x}^1, \vec{x}^2 \ldots \vec{x}^{|P|} >$, where, $| P |$ is the size of the population. For our task, binary representation of the solution is followed having length equals to the number of sentences present in the article. Each solution may have a varied number of sentences generated randomly between the range $[S_{min}, S_{max}]$. If the jth component of the solution is 1, then jth sentence should be part of summary and vice-versa. Solution representation is shown in Fig 1 assuming that article has 12 sentences.

[7]https://grobid.readthedocs.io/en/latest/Grobid-service/

4.3 Calculation of Objectives Functions

After initializing the population, objective functional values are computed for each solution, which help in evaluating the quality of the solution (or summary as the solution represents a summary). Note that the proposed framework is very generic and user can select any combination of objective functions.

4.4 Genetic Operators

For any evolutionary algorithm, in order to explore the search space efficiently or to find the global optimum solution by generating new solutions, various genetic operators are used which are mating pool generation, mutation, and crossover. Here also, new solutions/trial vectors are generated for every solution in each generation to form a new population, $P^{'}$. This step is shown by step-4 of the Figure-2. The process followed for new solution generation is described below. Let $\vec{x}_c$ be the current solution in the population for which new solution is to be generated.

4.4.1 Mating Pool Generation. The mating pool includes a set of solutions which can mate to generate new solutions. For the construction of the mating pool for the current solution, ($\vec{x}_c$), a fixed number of random solutions are picked up from the population.

4.4.2 Mutation and Crossover. Mutation is the change in component value of the solution, while, crossover is the exchange of component values between two solutions. In our work, we have used 2 trial vector generation schemes/variants namely, current-to-rand/1/bin and current-to-best/1/bin. These schemes have distinct properties. First one helps in creating diverse solution from the current solution (which further helps in introducing diversity among solutions), while, second one helps in speed up the convergence rate (provides right direction in reaching towards global optimal solution). Moreover, F and CR are two crucial parameters present in MBDE framework which help in generating good quality solutions or achieving faster convergence. In the literature [19, 37], the range of value suggested for F usually lies between 0.4 and 1, while for CR, value of 0.9 or 1 is suggested. But, sometime, fixing the values of these variables makes the search space limited. Therefore, instead of fixing them, pool of F and CR values are provided motivated by the paper [42] and discussed schemes can select these parameter values randomly from the given pools. Descriptions of these variants are provided in the paper [42] in continuous space. But, as our approach is based on binary encoding, therefore, they are adopted in binary space motivated by the paper [41]. To generate new trial vectors corresponding to $\vec{x}_c$, all schemes first make use of mutation and then crossover which are discussed below:

(1) *current-to-rand/1/bin:*

a) Mutation: To perform this operation for the current solution $\vec{x}_c$, firstly three random solutions, $\vec{x}_{r1}$, $\vec{x}_{r2}$, and, $\vec{x}_{r3}$, are selected from its constructed mating pool and then a probability vector $P(x^t)$ is generated by following the following operation:

$$P(x_j^t) = \frac{1}{1 + e^{-\frac{2\times b[x_{c,j}^t + r\times(x_{r1,j}^t - x_{c,j}^t) + F\times(x_{r2,j}^t - x_{r3,j}^t) - 0.5]}{1+2F}}} \tag{9}$$

where $\vec{x}_c$ is the current solution at generation 't' for which new solution is generated, $P(x_j^t)$ is the probability estimation operator, $(x_{c,j}^t + r\times(x_{r1,j}^t - x_{c,j}^t) + F\times(x_{r2,j}^t - x_{r3,j}^t) - 0.5)$ is the mutation operation, b is a real positive constant, r is a random number between 0 to 1, F is the DE control parameter, $\vec{x}_{k,j}$ is the jth component of kth solution for $k = \{r1, r2, r3, c\}$ at generation 't'. This operator generates probability value for each component of the current solution.

Then the corresponding offspring, y', for the current solution, $\vec{x}_c$ is generated as

$$y_j^{'} = \begin{cases} 1, \text{if rand}() \leq P(x_j^t) \\ 0, otherwise \end{cases} \tag{10}$$

where $j = 1, 2, \ldots, N$, N is the length of the solution and $rand()$ is a random number generated between 0 to 1. If random probability corresponding to a specific component of the current solution ($\vec{x}_c$) is less than the probability value generated for the same component using Eq. 9, then the value 1 will be assigned to a new solution at the same component; otherwise, 0 will be assigned.

b) Crossover: It is used for the exchange of components of mutated solution, y', and current solution, $\vec{x}_c$. After performing crossover, a new solution, y'', is generated, called as trail vector and is expressed as follows:

$$y_j'' = \begin{cases} y_j', \text{if rand}() \leq CR \\ x_j, \text{Otherwise} \end{cases} \tag{11}$$

where $rand()$ is a random probability between 0 to 1, $j = 1, 2, \ldots, N$, N is the length of the solution, CR is the crossover probability.

(2) *current-to-best/1/bin:* This variant makes use of two random solutions selected from the mating pool, current solution ($\vec{x}_c$) and the best solution $\vec{x}_{best}$ to generate a trial vector. Similar to current-to-rand/1/bin, it also first performs the mutation and then crossover. To select the best solution in the current generation, some mechanism like non-dominated sorting [8] can be used, but, it will increase computation time. Therefore, in our approach, the best solution is selected by considering the average of the used objectives functions (mathematically shown in Eq. 12).

$$f(\vec{x}_{best}) = \arg max_{i=1,2,\ldots,|P|} \Big(\sum_{j=1}^{m} Ob_{ij} \Big)/m \tag{12}$$

where $|P|$ and m are the size of the population (or number of solutions in the population) and the number of used objective functions, respectively, Ob_{ij} is the jth objective function value corresponding to ith solution. Then the following operation is performed to generate the probability vector which is further converted into binary space.

$$P(x_j^t) = \frac{1}{1 + e^{-\frac{2 \times b[x_{c,j}^t + r \times (x_{best,j}^t - x_{c,j}^t) + F \times (x_{r1,j}^t - x_{r2,j}^t) - 0.5]}{1+2F}}} \tag{13}$$

where $\vec{x}_{best}$ is the best solution at generation 't', $\vec{x}_c$ is the current solution at generation 't' for which new solution is generated. Rest of the notations are same as in current-to-rand/1/bin. Then Eqs. 10 and 11 are followed to generate the trial vector.

Out of two vectors, one having good objective function values will be considered as the best trail vector for the current solution [7]. To find the best trial vector, we have again used the concept of maximum average objective functional values.

Checking of Constraints: After application of mutation and crossover operations, constraint of number of 1s in the new solutions/trial vectors is checked. It may be possible that generated new solutions don't satisfy the constraint.

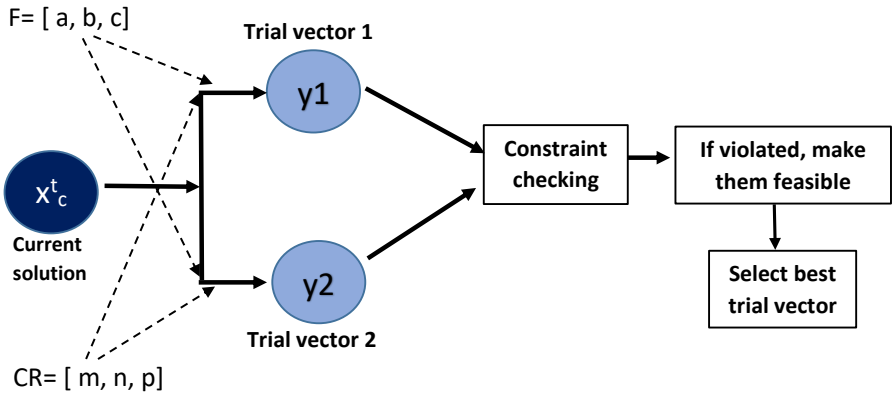


Fig. 3. Flow chart of generation of solutions from the current solution, $\vec{x}_c$ at generation 't' using two DE variants. Here, F and CR are the pool of some values; y1 and y2 are the trial vectors generated using current-to-rand/1/bin and current-to-best/1/bin scheme, respectively.

Therefore, to make them feasible (within constraint) some heuristics can be applied. The following steps are executed to make the new solutions feasible or within the range, $[S_{min}, S_{max}]$:

- Let us denote the new solution (y'') as ith solution
- Initialize $ModifiedSolution$ with zeros equal to the maximum length of the solution
- Sort the sentences present in the ith solution based on maximum number of uni-grams/maximum number of bi-grams/similarity with figure's caption. To select a single selection criterion, a random probability 'p' is generated. If $p < 0.33$ then sentences in the solutions are sorted based on maximum number of uni-grams; if $p > 0.33$ and $p < 0.67$, then sentences in the solutions are sorted based on maximum number of bi-grams; otherwise, those are sorted based on maximum similarity with figure's caption.
- Generate a random number 'r' between S_{min} and S_{max}.
- Fill the indices of $ModifiedSolution$ with 1s until we cover 'r' indices. Note that indices are considered in the sorted order as done in step-3.
- Return the $ModifiedSolution$.

The objective functional values of generated new solutions are also evaluated. The flow-chart of this entire process of solution generation is shown in Figure 3.

4.5 Selection of Best |P| Solutions for Next Generation

After forming a new population, P', it is merged with the old population, P. It is important to note that size of the population P' equals to the size of the population, P. Out of these merged solutions, only best $|P|$ solutions are selected using the dominance and non-dominance relationships between the solutions in the objective space. For this purpose, we have utilized the non-dominating sorting (NDS) and crowding distance based operators [8].

4.6 Termination Condition

The process of mating pool generation, crossover, and mutation followed by selection and then updation of the population is repeated until a maximum number of generations, g_{max} is reached. In other words, the loop will continue

until $g < g_{max}$. Here, g is the current generation number initialized to 0 and is incremented by 1 after each iteration. This step is shown by the diamond box in Figure 2.

4.7 Selection of Single Best Solution and Generation of Summary

After the final generation, we obtain a set of non-dominated solutions on the final Pareto optimal front. All these solutions are non-dominating to each other, thus, having equal importance. Therefore, the decision-maker has to select a solution based on his/her requirement. In this paper, for the purpose of reporting and comparative study, summary corresponding to each of the Pareto optimal solutions is generated and then, that solution is selected which has the highest F-measure value. In calculation of F-measure, it makes use of gold/reference summary. The sentences, in summary, are reported based on their occurrences in the scientific article. For example, the sentence which appears first in the article will be the first sentence in the summary. However, in real time, the reference summary may not be available. But, in this paper, the goal is to show that our proposed approach is able to generate a good summary for a given figure and by averaging results of best summaries of different figures, we are able to beat the existing algorithms.

5 EXPERIMENTAL SETUP

In the subsequent sections, we have discussed datasets, evaluation measures, and, parameters used.

5.1 Datasets

For our figure-summarization task, we have used two publicly available[8] data sets. First dataset, *FigSumGS1*, has 91 figures, while, second dataset, *FigSumGS2*, has 84 figures. Actual/gold summary is made available by the annotators. These figures belong to 19 biomedical full-text articles. Brief description of the used datasets in terms of the number of figures in each article, number of sentences in the article and in the gold summary of each figure etc., are provided in the supplementary sheet.

Table 3. Parameter setting for our proposed approach. Here, Q is the number of sentences in the actual summary specific to a figure.

Parameters	Values
Population size (|P|)	40
Maximum number of generations (g_{max})	25
F_{pool}	[0.6, 0.8, 1.0]
CR_{pool}	[0.1, 0.2, 1.0]
S_{min} and S_{max}	$Q + 2$ and $Q - 2$

5.2 Experimental Settings

Different parameter values used in our proposed framework are reported in the Table 3. Population size and maximum number of generations are kept fixed because more will be their values, more will be the computation time. Results obtained are averaged over 5 runs of the algorithm. For representation of sentences, BioBERT, a pre-trained model[9] on biomedical text articles and a book corpus were used which provide fixed length vectors of the sentences. To evaluate the performance of our system in comparison to available gold summary, we have reported the F-measure (or F1-score)

[8] http://figshare.com/articles/Figure_Associated_Text_Summarization_and_Evaluation/858903
[9] https://github.com/naver/biobert-pretrained/releases/tag/v1.0-pubmed-pmc

[25] value which is a well known measure in information retrieval. Formal definition of the F-measure is provided in the supplementary sheet.

5.3 Comparative Methods

As our proposed approach is unsupervised in nature, therefore, we have made comparison with other existing unsupervised methods. Although, supervised techniques exist in the literature, but, it will be unfair to make comparison between supervised and unsupervised methods. Unsupervised methods include three methods namely, Randomsent, FigSum [2], FigSum+ [25]. Further, three variants of FigSum+, which are similarity, tfidf, and, SurfaceCue based versions, are considered (shown in Table 6(a)). These variants select top-n sentences based on maximum caption similarity function, TF-IDF [26, 35] based similarity function, and, sentence referring to figure function, respectively. Here, TFIDF is a well known bag-of-words model in vector space. Brief descriptions of these methods are already provided in the related work section (Section 2). To the best of our knowledge, there is no other work in figure summarization after [25]. Note that our developed method is unsupervised in nature. Gold summaries were used only to evaluate our system at the end. Moreover, the system proposed is based on extraction of relevant sentences from the article related to a given figure; therefore, only sentence-extraction based methods are used for comparative study.

Table 4. Average precision (P), recall (R) and F-measure (F1) values obtained for both datasets using reduced set of sentences. Here, the decimal number in the left of '±'is the standard deviation.

		Datasets→	FigSumGS1			FigSumGS2		
S.No.	SAR version↓	Objective functions↓	P	R	F1	P	R	F1
1	SAR_CS	SRF	0.18±0.22	0.15±0.20	0.17±0.21	0.22±0.15	0.18±0.13	0.20±0.14
	SAR_TE		0.22±0.27	0.15±0.19	0.18±0.22	0.25±0.13	0.20±0.11	0.22±0.12
2	SAR_CS	STE+SRF	0.20±0.22	0.18±0.19	0.19±0.20	0.22±0.14	0.19±0.12	0.20±0.13
	SAR_TE		0.22±0.24	0.18±0.19	0.20±0.20	0.21±0.14	0.18±0.12	0.19±0.13
3	SAR_CS	STE+SOC1+SOC2	0.20±0.21	0.18±0.19	0.19±0.20	0.22±0.14	0.20±0.13	0.21±0.13
	SAR_TE		0.19±0.21	0.16±0.17	0.17±0.18	0.22±0.14	0.19±0.11	0.20±0.12
4	SAR_CS	SRF+SOC1+SOC2	0.19±0.21	0.18±0.20	0.18±0.20	0.22±0.13	0.20±0.13	0.21±0.13
	SAR_TE		0.21±0.25	0.17±0.21	0.18±0.22	0.21±0.13	0.18±0.12	0.20±0.12
5	SAR_CS	STE+SRF+SOC1+SOC2	0.21±0.22	0.19±0.21	0.20±0.21	0.21±0.22	0.17±0.21	0.18±0.21
	SAR_TE		0.23±0.25	0.19±0.21	**0.20±0.22**	0.24±0.14	0.20±0.12	**0.22±0.13**

6 EXPERIMENTAL RESULTS AND DISCUSSION

We have conducted two sets of experiments, ExpSet1 and ExpSet2, by varying the number of input sentences. We have discussed them one by one with corresponding results obtained. Then we have discussed the comparative analysis with the existing methods with ablation study on different combinations of objective functions. At the end, we have provided error analysis of the results obtained followed by statistical significance test of our results.

(1) *ExpSet1*: In this set, we have considered only those sentences in the article for our experiment whose entailment probability values to a given figure's caption (Let's say Fig-m is to be summarized) are greater than 0.5. The proposed approach is then applied on this reduced number of sentences. Note that the number of input sentences are reduced to minimize the computation time. This was done to see whether the reduced set of sentences extracted from the article using entailment probability values are sufficient to obtain a good quality summary.

Results and Discussion: The results obtained under *ExpSet1* are shown in Table 4. We have tried only 5 combinations of objective functions using different versions (SAR_CS and SAR_TE) of anti-redundancy objective function (SAR). From this Table, it can be observed that the highest values of F1-measure for FigSumGS1 and FigSumGS2 datasets are 0.21 and 0.22, respectively. These highest values are obtained using the SAR_TE in combination with objective functions, namely, STE, SRF, SOC1, and SOC2. In most of the rows in this table, the values of F-measure corresponding to SAR_TE are high. Thus, here we can infer that the anti-redundancy objective function measured in terms of textual entailment relationship is contributing towards the better result.

Table 5. Average precision (P), recall (R) and F-measure (F1) values obtained by the proposed approach for both datasets namely, FigSumGS1 and FigSumGS2, by varying the objective function combinations. Here, the decimal number in the left of '±'is the standard deviation. Note that here all sentences in the article are used for the experiment.

		Datasets→	FigSumGS1			FigSumGS2		
S.No.	SAR version↓	Objective functions↓	P	R	F1	P	R	F1
1	SAR_CS	STE	0.24±0.18	0.20±0.15	0.22±0.16	0.22±0.12	0.19±0.11	0.20±0.22
	SAR_TE		0.28±0.18	0.22±0.14	0.24±0.15	0.26±0.13	0.22±0.11	0.24±0.12
2	SAR_CS	SRF	0.53±0.17	0.46±0.20	0.49±0.17	0.31±0.13	0.27±0.12	0.29±0.12
	SAR_TE		0.64±0.24	0.47±0.18	0.54±0.19	0.39±0.14	0.30±0.11	0.34±0.12
3	SAR_CS	SFC	0.36±0.18	0.27±0.14	0.30±0.15	0.30±0.12	0.24±0.10	0.27±0.11
	SAR_TE		0.29±0.21	0.21±0.16	0.24±0.18	0.31±0.13	0.25±0.11	0.28±0.12
4	SAR_CS	STE+SRF	0.51±0.17	0.46±0.18	0.48±0.16	0.32±0.12	0.30±0.12	0.31±0.12
	SAR_TE		0.62±0.23	0.48±0.19	0.53±0.19	0.37±0.14	0.30±0.11	0.30±0.12
5	SAR_CS	STE+SFC	0.36±0.18	0.30±0.16	0.32±0.16	0.25±0.14	0.23±0.11	0.24±0.12
	SAR_TE		0.28±0.21	0.23±0.18	0.25±0.19	0.30±0.12	0.25±0.10	0.27±0.11
6	SAR_CS	SRF+SFC	0.54±0.17	0.47±0.18	0.50±0.16	0.34±0.13	0.28±0.11	0.31±0.12
	SAR_TE		0.63±0.21	0.47±0.16	0.53±0.17	0.37±0.14	0.30±0.12	0.33±0.12
7	SAR_CS	STE+SOC1+SOC2	0.43±0.17	0.42±0.18	0.42±0.17	0.32±0.13	0.30±0.13	0.31±0.12
	SAR_TE		0.50±0.20	0.44±0.20	0.46±0.19	0.37±0.13	0.31±0.11	0.34±0.37
8	SAR_CS	SRF+SOC1+SOC2	0.55±0.14	**0.54±0.18**	0.54±0.5	0.37±0.13	0.33±0.12	0.34±0.12
	SAR_TE		0.65±0.20	0.52±0.19	0.57±0.18	0.38±0.13	0.32±0.11	0.35±0.12
9	SAR_CS	STE+SRF+SOC1+SOC2	0.54±0.15	0.52±0.18	0.52±0.15	0.36±0.12	0.32±0.12	0.34±0.12
	SAR_TE		**0.65±0.20**	**0.54±0.18**	**0.59±0.18**	**0.42±0.12**	**0.38±0.11**	**0.40±0.11**

(2) *ExpSet2*: In this set, all the available sentences in the article are considered for our experiments. The proposed approach is applied on this full set of sentences.

Results and Discussion: The results obtained using all sentences of the articles are reported in Table 5. Further, in the same table, results are shown using different versions of anti-redundancy (SAR_CS and SAR_TE) objective function in combination with other objective functions. After observing Table 5, it is found that the highest values of F1-measure for FigSumGS1 and FigSumGS2 datasets are 0.59 and 0.40, respectively, which are more than values obtained after experimentation with reduced set of input sentences. Moreover, the maximum F-measure value obtained using different objective function combinations including SAR_CS function is 0.54 (S.No. 8) which is 4% less than the highest F-score. The other observations made from Table 5 are enumerated below:

(a) Among-st the most of the objective function combinations, SAR_TE performs better than SAR_CS. Thus, we can say that SAR_TE is contributing more in figure summarization process in comparison to SAR_CS.

(b) When we remove STE from the best combination (S.No. 9), F-score decreases by 1% (S.No. 8). But, on comparing, STE_TE+STE+SOC1+SOC2 (S.No. 7) and SRF_TE+STE+SOC1+SOC2 (S.No. 8), the second one is better. This

infers that although STE is contributing towards the best F-score value, SRF is more contributing than STE when used with SOC1 and SOC2. The same can also be observed by seeing the F-score of SAR_TE+STE (S.No. 1) and SAR_TE+SRF (S.No. 2). There is a big jump in the F-score value.

(c) On comparing, STE, SRF, and, SFC along with any version of SAR, again, SRF is more contributing. For any scientific article, it is purely logical because if a sentence refers to a particular figure keyword like 'Figure-<number>', then it indicates that sentence is associated with that figure.

Table 6. Comparison of the best results obtained by our proposed approach with (a) unsupervised methods; (b) supervised methods, in terms of average precision (P), recall (R) and F-measure (F1) for both datasets namely, FigSumGS1 and FigSumGS2. Here, the decimal number in the left of '±'is the standard deviation. Note that here all sentences in the article are used for the experiment.

		FigSumGS1			FigSumGS2		
Type of Methods	Method	P	R	F1	P	R	F1
Unsupervised	Proposed (**FigSum++**)	0.65±0.20	**0.54±0.18**	**0.59±0.18**	0.42±0.12	**0.38±0.11**	**0.40±0.11**
	RandomSent	0.06±0.09	0.06±0.12	0.06±0.09	0.08±0.08	0.09±0.11	0.08±0.09
	FigSum	0.28±0.24	0.19±0.19	0.22±0.19	0.31±0.20	0.13±0.10	0.18±0.13
	FigSum+ (SurfaceCue)	**0.96±0.13**	0.41±0.22	0.54±0.21	**0.63±0.36**	0.16±0.13	0.24±0.17
	FigSum+ (tfidf)	0.30±0.25	0.34±0.24	0.30±0.20	0.27±0.22	0.20±0.14	0.29±0.15
	FigSum+ (Similarity)	0.28±0.20	0.38±0.28	0.30±0.22	0.31±0.16	0.28±0.16	0.22±0.16

(a)

		FigSumGS1			FigSumGS2		
Type of Methods	Method	P	R	F1	P	R	F1
Unsupervised	Proposed (**FigSum++**)	**0.65±0.20**	0.54±0.18	**0.59±0.18**	0.42±0.12	0.38±0.11	**0.40±0.11**
Supervised	NBSurfaceCues	0.44±0.11	0.17±0.20	0.18±0.15	0.49±0.06	0.05±0.04	0.08±0.05
	NBSOTA	0.44±0.15	**0.74±0.17**	*0.53±0.12*	0.37±0.14	**0.43±0.19**	0.38±0.13
	SVMSOTA	0.58±0.15	0.17±0.20	0.23±0.22	**0.54±0.12**	0.10±0.11	0.15±0.15
	NBSimilarity	0.48±0.18	0.15±0.12	0.20±0.12	0.42±0.14	0.10±0.08	0.14±0.08

(b)

6.1 Comparison with Existing Unsupervised Methods

In Table 6(a), the best results obtained by our proposed approach in comparison to some existing unsupervised state-of-the-art techniques are shown. From this table, it can be observed that our proposed unsupervised method (FigSum++) attains the maximum F-measure values of 0.59 and 0.40 for FigSumGS1 and FigSumGS2 datasets, respectively, using combination of SAR_TE, STE, SRF, SOC1, and, SOC2, objectives functions (this result corresponds to the best result reported in Table 5). Although, for *FigSum+ (SurfaceCue)* method, Precision values are high (0.96 and 0.63 for two datasets), but, Recall (0.41 and 0.16) values are low as of our proposed method. It indicates that the number of sentences in the obtained summaries corresponding to this method are less and those are exactly matching to the sentences of the gold summaries. The technique, Randomsent, does not consider any feature specific objective function while generating summary. It randomly selects top-n sentences as a part of figure's summary and thus gives a very poor F-measure values of 0.06 and 0.08 for the used datasets, respectively. Note that our technique is based on sentence selection for figure summary. Therefore, we have made a comparison using only those techniques which also extract sentences for generating the summary. Out of three variants of FigSum+, SurfaceCue method gives F-measure values of 0.54 and 0.24 on the two datasets which are 5% and 16% less than the best values attained by our proposed unsupervised method. Note that we have not reported the number of sentences in the predicted summary corresponding to each figure as average F-measure values over all figures are reported in Table 6.

We have also compared our results in comparison to some supervised methods in Table 6(b). For comparison, we have considered different methods namely, NBSurfaceCue [25], NBSOTA [6], NBSimilarity [25], and, SVMSOTA [6]. Here, the first three methods (NBSurfaceCue, NBSOTA and NBSimilarity) make use of naive bayes classifier [27], while, fourth one (SVMSOTA) makes use of support vector machine [40]. The features used by SVMSOTA and NBSOTA to train the supervised model are sentence referring to figure, paragraph referring to figure, reference sentence similarity, caption similarity, etc. Although it is quite unfair to compare two different types of techniques (supervised and unsupervised) because in most of the cases supervised methods always perform better. But, here, after observing the results, it can be concluded that our F-measure value is better than existing supervised methods. In terms of improvements in F-measure values among supervised methods, we can say that there are 6% and 2% improvements obtained by our method for FigSumGS1 and FigSumGS2 datasets, respectively. But, recall value of NBSOTA is better than ours. This is because of using feature 'figure reference paragraph'while training, but, our system does not make use of any such paragraph-based feature. Pareto optimal solutions obtained after application of our proposed approach at the end of generation 0 and 24 are shown in the supplementary sheet due to length restriction.

6.2 Measure of Closeness of the Pareto Optimal Solutions in Different Runs of the Proposed Algorithm

At the end of the execution of our algorithm, we get a set of Pareto optimal solutions which may be different in various runs of the proposed algorithm. Therefore, to check the closeness of the Pareto optimal solutions in different runs of the algorithm, we have reported the generation distance (GD) [15]. It measures the convergence of the obtained Pareto optimal front (containing Pareto optimal solutions) by our proposed approach towards the true Pareto optimal front. Mathematically, it is defined as

$$GD = \frac{(\sum_{i=1}^{|Q|} d_i^p)^{\frac{1}{p}}}{|Q|} \quad \text{where} \quad d_i = \min_{i \in Q, j \in Q^*, k=1}^{|Q^*|} \sqrt{\sum_{m=1}^{M} |f_m^i - f_m^j|^2} \tag{14}$$

Where, Q and Q^* are the obtained and the actual Pareto optimal front, respectively, M is the number of objective functions, f_m^i is the mth objective function value of ith solution. In our summarization task, we don't know the actual Pareto optimal front; therefore, the Pareto optimal front obtained in the first run (Run1) of the algorithm is considered as the actual one. On the other hand, in other runs (Run2 and Run3) of the proposed algorithm, they are regarded as the obtained Pareto optimal fronts. Note that the best result was obtained when five objectives functions, SAR_TE, STE, SRF, SOC1, and SOC2, are optimized simultaneously, therefore, the Pareto optimal solutions corresponding to Run1, Run2 and Run3 are obtained by optimizing the same set of objective functions. Thus, here, the value of M is 5. Generally, the value of p is considered as 2, but, in this paper, we have considered it as 1.

We have randomly picked a total of 12 Figures from both the datasets and reported the results obtained in terms of GD in Table 7. In the Table, the left-hand side of the arrow ($\rightarrow$) indicates the actual Pareto optimal front, while, on the right-hand side, the obtained Pareto optimal front. From this Table, it can be seen that GDs of $Run1 \rightarrow Run2$ and $Run1 \rightarrow Run3$ are less than 0.1 which indicates that Pareto optimal solutions are almost same in every run of the proposed algorithm. Moreover, we have also reported their average, i.e., $[GD(Run1 \rightarrow Run2) + GD(Run1 \rightarrow Run3)]/2$, which is also less than 0.1.

6.3 Number of Fitness Function Evaluations

Generally, in any evolutionary based optimization strategy, the number of fitness function evaluations (NFE) [3, 34] is reported which equals to $g_{max} \times |P| \times M$, where, g_{max}, $|P|$ and M are the maximum number of generations, number of

Table 7. Closeness of the Pareto Optimal Solutions in different runs of the proposed algorithm using generational distance. Here, combination of the different objective functions corresponding to the best result shown in Table 5 is used for optimization; Run1, Run2 and Run3 indicate that we have executed our algorithm three times; $a \rightarrow b$ denotes the closeness of Pareto Optimal Solutions by Run-a with those by Run-b; second column indicates the figure number of the biomedical article shown with article number in the first column.

Article No.	Figure No.	Run1 → Run2	Run1 → Run3	Average
111020	3	0.0547	0.0553	0.0504
1134656	8	0.0592	0.0414	0.0503
1156890	8	0.0421	0.0436	0.0428
19673075	6	0.0278	0.0496	0.0387
20459090	2	0.0269	0.0253	0.0262
21183645	3	0.0255	0.0257	0.0256
22473769	2	0.0038	0.0044	0.0041
23041342	5	0.0188	0.0185	0.0186
21183645	2	0.0250	0.0400	0.0325
22473769	1	0.0291	0.0340	0.0315
23041342	4	0.0108	0.0118	0.0113
19673075	2	0.0446	0.0489	0.0467

solutions in the population and the number of used objective functions, respectively. In our approach, values of these variables are 40, 25, and 6, respectively. Thus, the value of NFE in our approach is 6000. But, we are not able to compare our obtained NFEs with existing state-of-the-art techniques as those are not based on evolutionary procedures.

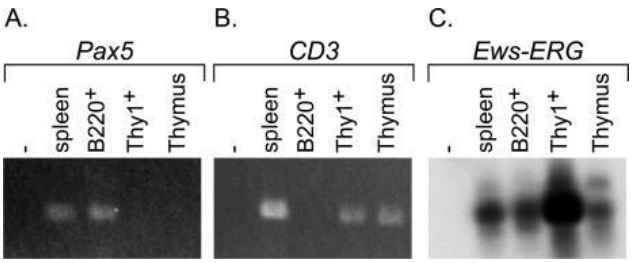
6.4 An Example of Summary Obtained

Here, we have shown an example of summary obtained by our proposed approach. The summary shown is corresponding to Figure-4 of the article available at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1159166 under *FigSumGS1* dataset and shown in Figure 4 of the current paper. Actual summary and figure's caption are also shown. The matching lines between actual and predicted summary are highlighted with the same colour. Note that the summary shown in Figure 4 is obtained after optimizing SAR_TE, SRF, SFC, SOC1, and, SOC2 objective functions. The F-measure value obtained corresponding to summary shown is 0.82, and, the number of sentences in the actual summary and predicted summaries are 9 and 8, respectively. This can be considered as an example of good summary as F-score is more than 80%.

6.5 Error Analysis

We have done a thorough error-analysis of the summaries generated for the figures in the articles with respect to both the data sets. This analysis is corresponding to the average best F-measure reported in Table 5 by our proposed approach.

6.5.1 For FigSumGS1 dataset: After observing the F-measure values for all figures in FigSumGS1 dataset, it has been found that only one figure has F-measure value less than 20% (Figure-3 of the article available at http://www.ncbi.nlm.nih.gov/pubmed/?term=22473769), 3 figures have F-measure values between 30% to 35%. For rest of the figures, the F-measure values are above 40%. The low value less than 20% is because of the following reason: the figure discuss the ratio of two biomedical terms, and, thus, the caption is full of only numbers, while, in the actual summary, sentences do n't have so many numbers. Our designed objective function mainly deals with the figure's caption at the syntactic



(a)

Caption: B and T Cells Express the Ews-ERG Fusion RNA A 96-d-old mouse with both Ews-ERG and Rag1-Cre alleles was used as a source of spleen and thymus cells. Single cell suspensions of spleen cells were labelled with anti-B220 or with anti-Thy1.2 and were purified using a MoFlo preparative flow cytometer. Estimated purities were achieved of greater than 95%. cDNA was prepared from RNA extracted from sorted cells or from aliquots of unsorted populations and RT-PCR (approximately 3,400 B220+ or 6,400 Thy1.2+ cell equivalents per PCR reaction) carried out with specific Pax5 (A), CD3 (B) or Ews-ERG (C) primers. PCR reaction products were fractionated on 1% agarose gels and either stained with ethidium bromide and photographed (A and B) or gel blotted and hybridised with an Ews-ERG probe (C)

(b)

Actual Summary	Predicted Summary
The possible inversion of the Ews-ERG gene in B cells was investigated using RT-PCR analysis of expressed Ews-ERG fusion mRNA Figure-4. RNA was prepared from whole spleen or thymus or from flow-sorted B220+ spleen cells 3 400 cells or Thy1.2+ spleen cells 6400 cells and RT-PCR performed. Pax5 and CD3 primers were used for specific detection of B cell and T cell transcripts respectively. Pax5 transcripts were detected in cDNA made from spleen and B220+ sorted cells and CD3 in the spleen thymus and thy1.2+ sorted cells Figure-4. The presence of Ews-ERG fusion RNA was analysed with RT-PCR primers yielding a product spanning the fusion junction that was detected with an internal junction probe. Ews-ERG RT-PCR product was detected in the unfractionated spleen and thymus sources as well as in the purified sorted B220+ and Thy1.2+ cells. Therefore, Cre-mediated inversion of the Ews-ERG gene occurs in both T and B cells. In this respect the absence of B cell tumours in the Ews-ERG invertors is of interest as both B and T cells undergo inversion of the Ews-ERG cassette see Figure-4 because Rag1-Cre is expressed in both cell types see Figure-S3. The absence of B cell tumours may reflect toxicity of the fusion protein for B cells although this seems unlikely given that we can detect the fusion mRNA in selected B220+ B lymphocytes see Figure-4 .	The possible inversion of the Ews-ERG gene in B cells was investigated using RT-PCR analysis of expressed Ews-ERG fusion mRNA Figure-4. RNA was prepared from whole spleen or thymus or from flow-sorted B220+ spleen cells 3 400 cells or Thy1.2+ spleen cells 6400 cells and RT-PCR performed. Pax5 and CD3 primers were used for specific detection of B cell and T cell transcripts respectively. Pax5 transcripts were detected in cDNA made from spleen and B220+ sorted cells and CD3 in the spleen thymus and thy1.2+ sorted cells Figure-4. The presence of Ews-ERG fusion RNA was analysed with RT-PCR primers yielding a product spanning the fusion junction that was detected with an internal junction probe. Therefore, Cre-mediated inversion of the Ews-ERG gene occurs in both T and B cells. PCR amplification from thymoma DNA was carried out with pools of Vb primers and a Jb2 reverse primer primer sequences from and sequences identified with the ImMunoGeneTics database. In this respect the absence of B cell tumours in the Ews-ERG invertors is of interest as both B and T cells undergo inversion of the Ews-ERG cassette see Figure-4 because Rag1-Cre is expressed in both cell types see Figure-S3.

(c)

Fig. 4. An example of Summary obtained by our proposed approach. (a) Figure-4 of the article available at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1159166; (b) Caption of the figure; (c) Actual and predicted summaries. Coloured lines (excluding black colour lines) in actual and predicted summary indicate the matched lines.

and semantic level, which tries to make our summary as close to caption and thus, there is little overlap between our summary and actual summary which decreases the F1 score value.

6.5.2 For FigSumGS2 dataset: In this dataset, there are mainly three figures (Figure-3, 5, and 6) of the article available at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1134656/ which have F1-scores, less than 20% and thus, causing decrease in overall average F1-score. Out of these, Figure-6 has F1 value of 0.09 which can be considered as an example of worst summary generated. This is due to the following reasons:

- The captions of these figures refer to caption of another figure (Figure-2 of the same article). The captions of Figure-3 and 6 have only 3 and 2 words, respectively, which are quite insufficient to explain a figure. For rest of the explanations, it is refers to caption of Figure-2.
- Second reason of very less F-measure specific to Figure-5 and Figure-7 of the same article is the following: some of the sentences (S) in the text refer to figures, but, the gold summary doesn't contain S. Also, inter annotator agreement for this dataset (*FigSumGS2*) is not available. This indicates some error in the annotation of gold summary.

6.6 Box-plots

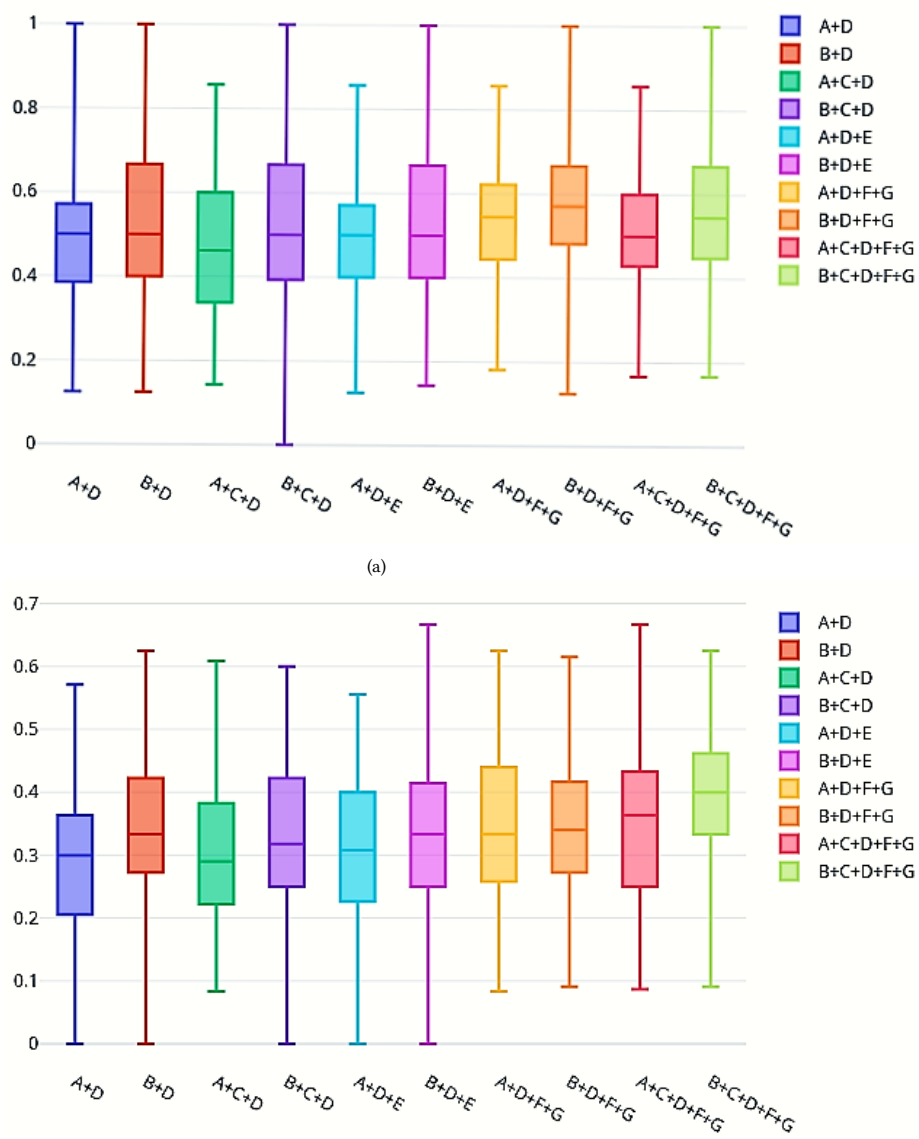
To illustrate the effectiveness of using SAR_TE over SAR_CS or in other words, to show the vaiations of F-measure values corresponding to two versions of the anti-redundancy objective function (SAR_CS and SAR_TE) in combination with other objective functions, we have drawn the box-plots for both the datasets. These box plots shown in Figure 5(a) and 5(b) correspond to FigSumGS1 and FigSumGS2 datasets, respectively. The results of five sets of objectives functions, i.e., SRF, STE+SRF, SRF+SFC, SRF+SOC1+SOC2, and, STE+SRF+SOC1+SOC2, each associated with SAR_CS and SAR_TE, are chosen for comparison because these combinations have equal or more than 50% and 30% F-measure values for FigSumGS1 and FigSumGS2 dataset, respectively. Thus, a total of 10 boxes are there in each figure. In each colored box, the horizontal colored line indicates the median value of F-measure mentioned at the y-axis. In these box-plots, the symbols namely, A, B, C, D, E, F, and, G represent SAR_CS, SAR_TE, STE, SRF, SFC, SOC1, and, SOC2, objective functions, respectively. From these plots, it can be analyzed that the median values of the used objectives functions in integration with SAR_TE, have high median value as a comparison to when used with SAR_CS. For example, the box corresponding to B+D (i.e., SAR_TE+SRF) has high median value than A+D (i.e., SAR_CS+SRF). Thus, it can be inferred that the anti-redundancy objective function measured in terms of textual entailment relationship is more effective than cosine similarity among-st sentences of the summary.

6.7 Statistical Significance of Results

To check the significance of our best result obtained with the existing state-of-the-art results (reported in Table 6), we have conducted the statistical significance t-test[10] at 5% significance level. This tests whether the best result obtained is statistically significant or occurred by chance. It provides p-value. Lesser is the p-value, more significant is our result. Note that there exist many papers on different applications of natural language processing like [4, 22, 30, 32–34] which use this significance level. Therefore, we have set the same level of significance in our approach. The p-values obtained using F-measure values reported in Table 6(a) are:

(1) .002695 for FigSumGS1 dataset

[10] https://www.socscistatistics.com/tests/studentttest/default2.aspx



(a)

(b)

Fig. 5. Box plots showing variations of the best F-measure values obtained for (a) FigSumGS1; (b) FigSumGS2 datasets. The symbols namely, A, B, C, D, E, F, and, G represent objective functions namely, SAR_CS, SAR_TE, STE, SRF, SFC, SOC1, and, SOC2, respectively.

(2) .000307 for FigSumGS2 dataset

Test results support the hypothesis that obtained improvements by the proposed approach are not occurred by chance, i.e., improvements are statistically significant.

6.8 Complexity Analysis of the Proposed Approach

In this section, we have analyzed the complexity of our proposed approach. Let the number of solutions, the number of objectives to be optimized and the maximum number of generations be N, M, and, g_{max}, respectively.

(1) Initialization of population takes $O(N)$ time as there are N solutions. For each solution, its objective functional values are calculated which takes $O(NM)$ time. Thus, the total time complexity of population initialization is $O(N + NM)$ which is equivalent to $O(NM)$.
(2) Construction of mating pool takes $O(1)$ time as solutions are randomly selected from the population.
(3) New solution generation using genetic operators (mutation and crossover) takes $O(2 \times (NM))$ time. The constant 2 is multiplied because for each solution, two trial vectors are generated i.e., total $2N$ new trial vectors will be generated and their associated objective function values are computed.
(4) Selection of best trail vector takes $O(1)$ time.
(5) Merging of old population (P) and new population ($P^{'}$) takes $O(1)$ time.
(6) Selection of the best solutions based on dominance and the non-dominance criteria from the merge population takes $O(M(2N)^2)$ time [8].

Steps-2 to 6 are repeated up to g_{max} number of generations. Note that step-2, 4 and 5, take constant time, therefore, they can be omitted from the total time complexity calculation. Thus, the total time complexity of the proposed architecture is

$$O(MN + g_{max}(2(NM) + M(2N)^2))$$

On solving further, it gives rise to

$$\implies O(MN + g_{max}(2NM + 4MN^2)) \equiv O(MN + g_{max}(4MN^2))$$
$$\implies O(MN(1 + 4g_{max}N)) \equiv O(4g_{max}MN^2))$$
$$\implies O(g_{max}MN^2))$$

which is the worst time complexity of our approach. From this comple≡xity, it can be inferred that if we increase the number of generations and the number of solutions in the population, then, there will be an increase in the computation time.

7 CONCLUSION AND FUTURE WORK

In this paper, we have proposed a sentence-based figure summarization system (FigSum++) for biomedical articles where relevant sentences relevant to a figure are extracted by optimizing different sentence scoring functions. These scoring functions include the semantic similarity with the caption, entailment to the figure's caption, number of sentences referring to figures, number of overlapping words between sentences and figure's caption, the dissimilarity between sentences (to remove redundancy from the summary), etc. and those are simultaneously optimized using multi-objective binary differential evolutionary (DE) algorithm. For efficient search or to reach towards global optimal solution, ensemble of two different DE variants is used in the proposed framework. Moreover, another function of measuring anti-redundancy in summary in terms of textual entailment is also proposed. To measure the semantic similarity among-st sentences, recently proposed, BioBERT language model for biomedical text mining is utilized. From the obtained results, it is evident that newly proposed anti-redundancy based objective function when measured in terms of textual entailment (TE) and optimized with other objective functions provides improvements of 5% and 11% for two datasets in terms of the F1-score over the state-of-the-art methods, respectively. Moreover, TE based

anti-redundancy objective function performs better than cosine similarity based anti-redundancy objective function. Thus, it can be inferred that textual entailment plays a major role in summarization task. Existing algorithms provide a single summary to the end-user, but, our approach provides varieties of summaries to the end-user (each varying in length and quality) and the user can select any summary based on his/her choice.

In the future, we would like to extend the proposed summarization system at the paragraph level. Instead of sentences, paragraphs referring to the figures will be automatically extracted. We would also like to parallelize our summarization system by simultaneously generating summaries of all the figures of a given article.

ACKNOWLEDGMENTS

Dr. Sriparna Saha gratefully acknowledges the Young Faculty Research Fellowship (YFRF) Award, supported by Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia) for carrying out this research.

REFERENCES

[1] Stergos Afantenos, Vangelis Karkaletsis, and Panagiotis Stamatopoulos. 2005. Summarization from medical documents: a survey. *Artificial intelligence in medicine* 33, 2 (2005), 157–177.

[2] Shashank Agarwal and Hong Yu. 2009. FigSum: automatically generating structured text summaries for figures in biomedical literature. In *AMIA Annual Symposium Proceedings*, Vol. 2009. American Medical Informatics Association, 6.

[3] Rasim M Alguliev, Ramiz M Aliguliyev, and Nijat R Isazade. 2012. DESAMC+ DocSum: Differential evolution with self-adaptive mutation and crossover parameters for multi-document summarization. *Knowledge-Based Systems* 36 (2012), 21–38.

[4] Rasim M. Alguliyev, Ramiz M. Aliguliyev, and Nijat R. Isazade. 2013. Multiple documents summarization based on evolutionary optimization algorithm. *Expert Syst. Appl.* 40 (2013), 1675–1689.

[5] Sanghamitra Bandyopadhyay, Sriparna Saha, Ujjwal Maulik, and Kalyanmoy Deb. 2008. A simulated annealing-based multiobjective optimization algorithm: AMOSA. *IEEE transactions on evolutionary computation* 12, 3 (2008), 269–283.

[6] Sumit Bhatia and Prasenjit Mitra. 2012. Summarizing figures, tables, and algorithms in scientific publications to augment search results. *ACM Transactions on Information Systems (TOIS)* 30, 1 (2012), 3.

[7] Swagatam Das and Ponnuthurai Nagaratnam Suganthan. 2011. Differential evolution: A survey of the state-of-the-art. *IEEE transactions on evolutionary computation* 15, 1 (2011), 4–31.

[8] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation* 6, 2 (2002), 182–197.

[9] Daniel M Dunlavy, Dianne P OâĂŹLeary, John M Conroy, and Judith D Schlesinger. 2007. QCS: A system for querying, clustering and summarizing documents. *Information processing & management* 43, 6 (2007), 1588–1605.

[10] Soumi Dutta, Vibhash Chandra, Kanav Mehra, Asit Kumar Das, Tanmoy Chakraborty, and Saptarshi Ghosh. 2018. Ensemble Algorithms for Microblog Summarization. *IEEE Intelligent Systems* 33, 3 (2018), 4–14.

[11] Rafael Ferreira, Luciano de Souza Cabral, Rafael Dueire Lins, Gabriel Pereira e Silva, Fred Freitas, George DC Cavalcanti, Rinaldo Lima, Steven J Simske, and Luciano Favaro. 2013. Assessing sentence scoring techniques for extractive text summarization. *Expert systems with applications* 40, 14 (2013), 5755–5764.

[12] Robert P Futrelle. 1999. Summarization of diagrams in documents. *Advances in Automated Text Summarization* (1999), 403–421.

[13] Robert P Futrelle. 2004. Handling figures in document summarization. *In Proceedings of the Workshop at the Annual Meeting of the Association for Computational Linguistics* (2004), 61–65.

[14] Eugene J Guglielmo and Neil C Rowe. 1996. Natural-language retrieval of images based on descriptive captions. *ACM Transactions on Information Systems (TOIS)* 14, 3 (1996), 237–267.

[15] Zhenan He and Gary G Yen. 2016. Visualization and performance metric in many-objective optimization. *IEEE Transactions on Evolutionary Computation* 20, 3 (2016), 386–402.

[16] Anna Huang. 2008. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, Vol. 4. 9–56.

[17] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746* (2019).

[18] Inderjeet Mani and Mark T Maybury. 2001. Automatic summarization. (2001).

[19] Ali Wagdy Mohamed, Hegazy Zaher Sabry, and Tareq Abd-Elaziz. 2013. Real parameter optimization by an effective differential evolution algorithm. *Egyptian Informatics Journal* 14, 1 (2013), 37–53.

[20] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[21] Shashi Narayan, Nikos Papasarantopoulos, Shay B Cohen, and Mirella Lapata. 2017. Neural extractive summarization with side information. *arXiv preprint arXiv:1704.04530* (2017).

[22] Tadashi Nomoto and Yuji Matsumoto. 2001. A new approach to unsupervised text summarization. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 26–34.

[23] Hilário Oliveira, Rafael Dueire Lins, Rinaldo Lima, and Fred Freitas. 2017. A regression-based approach using integer linear programming for single-document summarization. In *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 270–277.

[24] Rebecca Passonneau, Karen Kukich, Jacques Robin, Vasileios Hatzivassiloglou, Larry Lefkowitz, and Hongyan Jing. 1996. Generating summaries of workflow diagrams. In *Proceedings of the International Conference on Natural Language Processing and Industrial Applications*. 204–210.

[25] Balaji Polepalli Ramesh, Ricky J Sethi, and Hong Yu. 2015. Figure-associated text summarization and evaluation. *PloS one* 10, 2 (2015), e0115671.

[26] Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, Vol. 242. 133–142.

[27] Irina Rish et al. 2001. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, Vol. 3. 41–46.

[28] Alexey Romanov and Chaitanya Shivade. 2018. Lessons from Natural Language Inference in the Clinical Domain. *arXiv preprint arXiv:1808.06752* (2018).

[29] Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685* (2015).

[30] Sriparna Saha, Sayantan Mitra, and Stefan Kramer. 2018. Exploring multiobjective optimization for multiview clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 12, 4 (2018), 44.

[31] Naveen Saini, Sriparna Saha, and Pushpak Bhattacharyya. 2018. Automatic Scientific Document Clustering Using Self-organized Multi-objective Differential Evolution. *Cognitive Computation* (19 Dec 2018). https://doi.org/10.1007/s12559-018-9611-8

[32] Naveen Saini, Sriparna Saha, and Pushpak Bhattacharyya. 2019. Automatic Scientific Document Clustering Using Self-organized Multi-objective Differential Evolution. *Cognitive Computation* 11, 2 (2019), 271–293.

[33] Naveen Saini, Sriparna Saha, Aditya Harsh, and Pushpak Bhattacharyya. 2019. Sophisticated SOM based genetic operators in multi-objective clustering framework. *Applied Intelligence* 49, 5 (2019), 1803–1822.

[34] Naveen Saini, Sriparna Saha, Anubhav Jangra, and Pushpak Bhattacharyya. 2019. Extractive single document summarization using multi-objective optimization: Exploring self-organized differential evolution, grey wolf optimizer and water cycle algorithm. *Knowledge-Based Systems* 164 (2019), 45–67.

[35] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management* 24, 5 (1988), 513–523.

[36] Jesus M Sanchez-Gomez, Miguel A Vega-Rodríguez, and Carlos J Pérez. 2018. Extractive multi-document text summarization using a multi-objective artificial bee colony optimization approach. *Knowledge-Based Systems* 159 (2018), 1–8.

[37] Rainer Storn and Kenneth Price. 1997. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization* 11, 4 (1997), 341–359.

[38] Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. 2007. Manifold-Ranking Based Topic-Focused Multi-Document Summarization.. In *IJCAI*, Vol. 7. 2903–2908.

[39] Bing-Chuan Wang, Han-Xiong Li, Jia-Peng Li, and Yong Wang. 2018. Composite differential evolution for constrained evolutionary optimization. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 99 (2018), 1–14.

[40] Lipo Wang. 2005. *Support vector machines: theory and applications*. Vol. 177. Springer Science & Business Media.

[41] Ling Wang, Xiping Fu, Muhammad Ilyas Menhas, and Minrui Fei. 2010. A modified binary differential evolution algorithm. In *Life System Modeling and Intelligent Computing*. Springer, 49–57.

[42] Yong Wang, Zixing Cai, and Qingfu Zhang. 2011. Differential evolution with composite trial vector generation strategies and control parameters. *IEEE Transactions on Evolutionary Computation* 15, 1 (2011), 55–66.

[43] Peng Wu and Sandra Carberry. 2011. Toward extractive summarization of multimodal documents. In *Proceedings of the Workshop on Text Summarization at the Canadian Conference on Artificial Intelligence*. 53–61.

[44] Jen-Yuan Yeh, Hao-Ren Ke, Wei-Pang Yang, and I-Heng Meng. 2005. Text summarization using a trainable summarizer and latent semantic analysis. *Information processing & management* 41, 1 (2005), 75–95.

[45] Hong Yu, Shashank Agarwal, Mark Johnston, and Aaron Cohen. 2009. Are figure legends sufficient? Evaluating the contribution of associated text to biomedical figure comprehension. *Journal of biomedical discovery and collaboration* 4, 1 (2009), 1.

[46] Dan Zhang and Bin Wei. 2014. Comparison between differential evolution and particle swarm optimization algorithms. In *Mechatronics and Automation (ICMA), 2014 IEEE International Conference on*. IEEE, 239–244.