Multi-objective Based Approach for Microblog Summarization

Naveen Saini, *Student Member, IEEE,* Sriparna Saha, *Senior Member, IEEE,* and Pushpak Bhattacharyya

***Abstract*—In recent years, social networking sites such as Twitter have become the primary sources for real-time information of ongoing events such as political rallies, natural disasters etc. At the time of occurrence of natural disasters, it has been seen that relevant information collected from tweets can help in different ways. Therefore, there is a need to develop an automated microblog/tweet summarization system to automatically select relevant tweets. In this paper, we employ the concepts of multi-objective optimization in microblog summarization to produce good quality summaries. Different statistical quality measures namely, length, tf-idf score of the tweets, anti-redundancy, measuring different aspects of summary, are optimized simultaneously using the search capability of a multi-objective differential evolution technique. Different types of genetic operators including recently developed self-organizing map (a type of neural network) based operator, are explored in the proposed framework. To measure the similarity between tweets, word mover distance is utilized which is capable of capturing the semantic similarity between tweets. For evaluation, four benchmark datasets related to disaster events are used, and the results obtained are compared with various state-of-the-art techniques using ROUGE measures. It has been found that our algorithm improves by 62.37% and 5.65% in terms of ROUGE$-$2 and ROUGE$-$L scores, respectively, over the state-of-the-art techniques. Results are also validated using statistical significance t-test. At the end of the paper, extension of proposed approach to solve the multi-document summarization task is also illustrated.**

***Index Terms*—Microblog, disaster events, extractive summarization, multi-objective optimization, evolutionary algorithm, word mover distance.**

I. Introduction

Nowadays, social networking sites such as Twitter have become the main source for gathering real-time information of ongoing events such as political issue, man-made and natural disasters etc. In the literature [1], [2] importance of accessing microblogging sites for gathering information is shown. Vast amount of tweets are posted every-day and this makes the relevant information extraction from such data a cumbersome process. Moreover, it has been seen that common people stay connected with each other through microblogging sites at the time of occurrence of natural disasters and several useful information can be extracted from such tweets which can further help in managing the situation by the Govt. Therefore, there is a need to develop an automated microblog/tweet summarization [3], [4] system in which relevant tweets are selected automatically for extracting information based on various tweet scoring features.

All authors are with Department of Computer Science and Engineering, Institute of Technology Patna, Bihar, India-801103 .

E-mail:{naveen.pcs16, sriparna, pb}@iitp.ac.in

A. Motivation

Some existing tweet summarization techniques include cluster-rank [5], Lex Rank [6], LSA [7], Luhn [8], MEAD [9], SumBasic [10], SumDSDR [11] and COWTS [3]. A detailed descriptions of these algorithms is available at [12], [13]. The main drawback of these algorithms is that they consider a single statistical feature to assign a score to each tweet. For example, in COWTS, tweets' score is awarded based on coverage of important content words like nouns, verbs and numerals. But, there can be many relevant tweets which are important with respect to some other perspectives/themes like sum of tf-idf [14] scores of different words in the tweet etc. Motivated by this, in this paper, a novel microblog/tweet summarization technique (MOOTweetSumm) is proposed using the concepts of multi-objective optimization (MOO). Several tweet scoring features/objective functions like length of the tweet [4], tweet having maximum tf-idf score [4] are simultaneously optimized using the multi-objective binary differential evolution algorithm (MOBDE) [15] which is an evolutionary algorithm (EA) [16]. However, other optimization strategies like AMOSA [17], NSGA-II [18], etc. can also be used.

MOBDE consists of a set of solutions/chromosomes called as population. Each solution is represented as the binary vector denoting a set of possible tweets to be selected in the generated summary. Existing literature [19], [20], [21], [22] has shown that differential evolution algorithm (DE) performs better and faster compared to existing EAs in solving different optimization problems. This motivates us to use DE framework in our approach to optimize the objective functions.

As there can be lot of re-tweets, therefore, to avoid having redundant information as a part of the summary, another objective function namely, anti-redundancy is also optimized.

In recent years, self-organizing map (SOM) [23] based reproduction operator becomes popular in solving different tasks like document summarization [22], document clustering [24], etc. SOM is a type of neural network which maps high dimensional input space to low-dimensional output space, where, output space is a grid of neurons arranged in 2-dimensional space. The central principle behind the SOM is that input samples which are close to each other in the input space, should also come close to each other in the output space. Thus, it can be used as a cluster analysis tool. It has already been established that SOM based reproduction operators work good in the evolutionary algorithm [25], [26]. It was proved that these operators play vital roles in generating good quality new solutions from the old solutions over the generations and thus, help in reaching towards the global optimum solutions. Therefore, to establish whether SOM-based operator is effective in microblog summarization task or not, this is utilized in MOBDE framework. In our approach, SOM is first trained using the current population to discover the localities of chromosomes, and then a mating pool is constructed for each chromosome using the neighborhood relationships extracted by SOM. After that, chromosomes present in the mating pool are combined using reproduction operators (crossover and mutation) [27] to generate some new solutions. In addition to SOM-based operator, we have also explored normal DE operators (called as without SOM-based operator) in our proposed framework. It means that instead of using neighborhood relationships identified by the SOM, solutions are randomly selected from the population to construct the mating pool.

To measure the similarity/dis-similarity between tweets, recently proposed word mover distance (WMD) [28] is utilized. It is able to capture the semantic similarity between tweets.

Proposed approach is evaluated on four disaster event related datasets. Results obtained clearly show the superiority of our proposed algorithm in comparison to various state-of-the-art techniques. At a part of this paper, potentiality of the proposed approach is tested for multi-document summarization where we have to summarize a given set of documents.

B. Contribution

The major contributions of the current paper are enumerated below:

- A multi-objective optimization based approach is proposed for microblog summarization task in which different goodness measures of a summary are optimized simultaneously. As per the literature survey, it is the first attempt in using MOO framework for solving the microblog summarization task.
- Ablation study is presented to illustrate which combination of a set of objective functions is best suited for summarizing each dataset.
- Existing algorithms provide single summary after the execution of the algorithm. But, proposed approach outputs different possible summaries having variable number of tweets corresponding to different non-dominated solutions of the final Pareto front to the user. Therefore, the user will have more alternatives in selecting a single summary from the final pool. Depending on the user/domain requirement a single summary can be selected.
- The extension of the proposed approach is shown for multi-document summarization task.

The rest of the paper is organized as follows: Section II discusses the literature survey and related background knowledge. Section III discusses problem definition and its mathematical formulation. Section IV discusses the proposed approach. Experimental setup and result discussions are presented in Sections V and VI, respectively. In Section VII, the extension of proposed approach is shown for multi-document summarization task. Finally, the paper is concluded in Section VIII.

II. Related Works and Background Knowledge

In the literature, a lot of works have been done on tweet summarization. In [29], the problem of summarization of tweets related to sports event was solved. But, summarization of disaster event related tweets is more important as it may convey relevant information to the higher authorities and help them to take the desired action. In [30], first clustering of tweets is performed and some representative tweets are selected from each cluster. Then arrangement of these tweets is carried out using graph based LexRank [6] algorithm. Dutta et al. [12] showed the comparison among various extractive summarization techniques to summarize disaster related tweets. These techniques include Cluster-rank [5], Lex Rank [6], LSA [7], Luhn [8], MEAD [9], SumBasic [10], SumDSDR [11] and COWTS [3].

COWTS technique uses the content words of the tweets to generate the summary of situational tweets. Situational tweets are those tweets which provide information like status update, i.e., current situation in the effected region by the disaster event. The extension of COWTS work was done in [31]. In [32], time aware knowledge is extracted from the tweets for microblog summarization task. Recently, [4] proposed an ensemble approach for microblogs summarization which generates the summary after considering the summaries of various algorithms discussed in [12]. But, in real time, application of ensemble approach for summarizing tweets is time consuming as firstly, we have to generate the summaries by different algorithms and then produce the final summary by considering these individual summaries.

The paper by [33] works on identifying sub-event and summarizing disaster tweets. There also exist some works on Twitter data in a post-disaster scenario. Rudra et al. [34] developed a classifier to distinguish communal tweets during a disaster event. In [35], rumor identification schemes are developed based on the user's behaviour. In [36], a technique was developed to detect whether a tweet is a spam or not. But, these papers [3], [35], [36] don't discuss about summarization. Although, techniques, proposed in these papers may be useful before applying any summarization methods.

A. Multi-objective Optimization (MOO)

Multi-objective optimization (MOO) is a framework of optimizing more than one objective functions simultaneously and providing a set of alternative solutions known as Pareto optimal set to the decision maker. In other words, MOO problem can be formulated as

$$\max\{f_1(\vec{x}), f_2(\vec{x})...f_m(\vec{x})\} \quad \text{such that} \quad \vec{x} \in X \tag{1}$$

where $X = \{\vec{x_1}, \vec{x_2}...\vec{x_n}\}$ is a possible set of decision vectors in n-dimensional space, $m \geq 2$ and it denotes the number of objective functions to be maximized. Some constraints can also be a part of the optimization process.

B. Tweet Similarity/dis-similarity Measure: Word Mover Distance

Word Mover Distance (WMD) [28] calculates the dis-similarity between two texts as the amount of distance that the *embedded words [37] of one text needs to travel to reach the embedded words of another text [28].* In our approach, text means a tweet. To obtain word embeddings of different words, it makes use of word2vec [37] model. If two sentences are similar, then the corresponding WMD will be 0.

III. Problem Definition

Consider an event D consisting of N tweets, D=$\{t_1, t_2, \ldots, t_N\}$. Our main task is to find a subset of tweets, $T \subseteq D$, such that

$$S_{min} \leq \sum_{i=1}^{N} B_i \leq S_{max} \quad and \quad B_i = \begin{cases} 1, & \text{if } t_i \in T \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

$$\text{such that} \quad maximize\{Ob_1(T), Ob_2(T), Ob_3(T)\} \tag{3}$$

where, S_{min} and S_{max} are the minimum and the maximum number of tweets in the summary, respectively, Ob_1, Ob_2 and Ob_3 are the objective functions discussed in subsequent sections. Note that in Eq. 3, there can also be two objective functions ((Ob_1 and Ob_2) or (Ob_1 and Ob_3)) instead of three. These objective functions quantify the goodness of different tweets and further help in improving the quality of generated summary. All these objective functions have to be maximized simultaneously by the use of some multi-objective optimization framework. These objectives are calculated for each solution in the population as each solution denotes subset of tweets representing a summary.

1) Anti-redundancy (Ob_1): In a set of tweets, lot of re-tweets can be there, therefore, to reduce the redundancy in the summary, this objective function is considered. It is expressed as:

$$Ob_1 = \frac{\sum_{i,j=1, i \neq j}^{|T|} dist_{wmd}(t_i, t_j)}{|T|} \tag{4}$$

where, t_i and t_j are the ith and jth tweets, respectively belonging to T, $|T|$ is the total number of tweets to be in the summary, $dist_{wmd}(t_i, t_j)$ is the Word Mover Distance (for definition refer to section II-B) between ith and jth tweets.

2) Maximum tf-idf Score of the Tweets (Ob_2): tf-idf [14] is a well known measure in information retrieval to assign some weights to different words. Here, 'tf'' means term frequency and 'idf' means inverse-document frequency in a set of tweets (considered as a document). Each tweet is considered as a bag of words, each word having it's own tf-idf score. Thus, a tweet 't'can be represented as a vector

$$v_t = [w_{1t}, w_{2t}, w_{3t}, \ldots\ldots, w_{nt}] \tag{5}$$

where

$$w_{k,t} = tf_{k,t}.\left(1 + \log \frac{1+N}{1 + \{t' \in D | k \in t'\}}\right) \tag{6}$$

and '$tf_{k,t}$' is calculated by counting the number of occurrences of kth word in the same tweet (t), $t' \in D$, N is the total number of tweets available. Thus, summation of tf-idf scores of different tweets belonging to T is considered. The subset of tweets having maximum average tf-idf score is considered as a good summary. Mathematically, it can be expressed as

$$Ob_2 = \frac{\sum_{i=1}^{|T|} \sum_{word_k \in t_i, t_i \in T} w_{k,t_i}}{|T|} \tag{7}$$

where w_{k,t_i} is the tf-idf score of kth word ($word_k$) present in a tweet t_i and t_i is the ith tweet belonging to T.

3) Maximum length of the tweets (Ob_3): Based on the assumption that longer tweet conveys important information, this objective function is taken into consideration. Mathematically, it can be expressed as

$$Ob_3 = \sum_{i=1}^{|T|} length(t_i) \tag{8}$$

where, t_i is the ith tweet in the summary, $length$ counts the number of words in the tweet after removing stop words (example: is, am are etc.). However, some of the longer tweets may not be relevant as they contain irrelevant words. Therefore, other objective function discussed above (Ob_2) is considered which pays attention to the importance of different words in the tweet.

IV. Proposed Methodology

In this paper, we have developed an extractive tweet summarization system. It utilizes a multi-objective based differential evolution technique as the underlying optimization strategy. SOM-based genetic operators are incorporated in the process to see their effectiveness. The flowchart of the proposed approach is shown in Figure 1.

A. Representation of Solution and Population Initialization

Any evolutionary algorithm works with a population of solutions and population P consists of solutions (or chromosomes, both can be used interchangeably) $< \vec{x}^1, \vec{x}^2 \ldots \vec{x}^{|P|} >$. Each solution is represented as a binary vector. If a dataset or event has N tweets $\{t_1, t_2, \ldots, t_N\}$, then solution length will be N. For example, if an event consists of 10 tweets then a valid solution can be represented as $[1, 0, 0, 1, 1, 0, 1, 0, 0, 0]$. This solution indicates that first, fourth, fifth and seventh tweets of the original event are in the summary. Initial population is generated randomly having varied number of tweets between $[S_{min}, S_{max}]$. This provides the end user a flexibility to choose the best summary as per his/her requirement or expert knowledge in terms of the number of tweets.

B. Objective Functions Used

To obtain a good summary, use of good set of objective functions/quality measures is essential. These objective functions quantify the quality of the subset of tweets present in the solutions and thus optimization of all these helps in achieving good quality summary. All these objective functions are already discussed in section III and all are of maximization type.

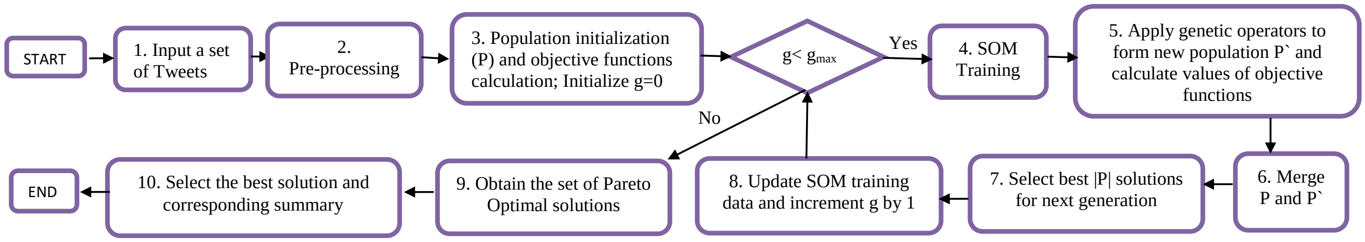


Fig. 1. Flow chart of the proposed architecture where, g is the current generation number initialized to 0 and g_{max} is the user-defined maximum number of generations (termination condition), $|P|$ is the size of population.

C. SOM Training

In this step, SOM training is performed using the solutions in the population. SOM [23], [38] is a special type of two-layered neural network and dimensionality reduction tool which maps high-dimensional data in input space to a low-dimensional space (output space). Usually, low-dimensional space consists of $2-$D grid of neurons. Each neuron is associated with two vectors: position vector and the weight vector. Position vector identifies the position of a neuron the in $2-$D, while, weight vector signifies the connection weights between the input vector and the neuron. The principle of SOM suggests that similar input patterns nearby to each other in the input space come close to each other in the output space. When an input pattern is presented to the grid, firstly winning neuron is determined using the shortest Euclidean distance criteria. Then the weight of the winning neuron and neighboring neurons (around the winning neuron) are updated so that these neurons form a region and become sensitive to the same type of input patterns. More details about the learning algorithm for SOM can be found in the paper [25]. Thus, SOM will help in understanding the distribution structure of the solutions in the population. In other words, SOM provides topology preserving map of the solutions in low dimensional space.

D. Genetic Operators

In our framework, from each solution, a new solution is generated using three steps: mating pool generation, crossover, and mutation and thus set of these new solutions form a new population (P$^{'}$). These genetic operators are described below:

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, only neighborhood solutions identified by the SOM are considered. Exploration and exploitation behaviors are also considered while generating the mating pool. Steps used to create it are described below. Let $\vec{x}_{current}$ be the current solution for which we want to generate a new solution, and β is some threshold probability.

1) Find the winning neuron 'b' in the SOM grid for $\vec{x}_{current}$ using the shortest Euclidean distance as $b = \arg min_{1 \le u \le U} \parallel \vec{x}_{current} - \vec{w}^u \parallel$, where $\vec{w}^u$ is the weight vector of u^{th} neuron, U is the total number of neurons.
2) Find the solutions $\in$ population, mapped to neighboring neurons around the winning neuron. These neighboring neurons are found out by calculating the Euclidean distances between the position vector of neuron 'b' and other neurons's position vectors.
3) Generate some random probability r.
4) If $r < \beta$, then after calculating the Euclidean distances between the winning neuron and others, neuron indices are sorted based on minimum distance to 'b'. Then solutions mapped to a fixed number of sorted neuron indices are extracted as we use fixed mating pool size in our approach. This step helps in exploiting the search space.
5) If $r > \beta$, then all solutions in the population will be part of the mating pool which helps in the exploration of the search space.

2) Mutation: To perform mutation on the current solution $\vec{x}_{current}$, firstly three solutions, $\vec{x}_{r1}$, $\vec{x}_{r2}$ and $\vec{x}_{r3}$ are selected randomly from it's mating pool and then following operation is performed

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

where $P(x_j^t)$ is the probability estimation operator, $(x^t_{r1,j} + F \times (x^t_{r2,j} - x^t_{r3,j}) - 0.5)$ is the mutation operation, b is a real positive constant, F is the scaling factor and $x^t_{r1,j}$, $x^t_{r2,j}$ and $x^t_{r3,j}$ are the jth components of randomly chosen solutions at generation 't'. Then the corresponding offspring y' for the current solution, $\vec{x}_{current}$ is generated as

$$y'_j = \begin{cases} 1, \text{ if rand()} \le P(x_j^t) \\ 0, otherwise \end{cases}$$

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

3) Crossover: Crossover operation is used for the exchange of components of the current solution, $\vec{x}_{current}$, and mutated solution, y' generated in section IV-D2. After crossover, it gives rise to new solution, y'', and it is expressed by the following equation:

$$y''_j = \begin{cases} y'_j, \text{if rand()} \le CR \\ x_j, \text{Otherwise} \end{cases}$$

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

1) Pick up any random solution. Let us call it ith solution.
2) Initialize $ModifiedSolution$ with zeros equal to the maximum solution length.
3) Find the indices of sorted tweets of the ith solution based on maximum tweet length or maximum tf-idf score. To do this, a random probability 'p' is generated. If $p < 0.5$ then solutions are sorted based on maximum tweet length, otherwise, those are sorted based on maximum tf-idf scores.
4) Generate a random number 'r' between S_{min} and S_{max}
5) Fill the indices of $ModifiedSolution$ with 1s until we cover 'r' indices. Note that filled indices are sorted indices obtained in step-3.
6) return the $ModifiedSolution$.

Here, it is important to note that while optimizing two objectives, Ob_1 and Ob_2, we also provide importance to new solution generated using maximum tweet length (based on some probability as in step-3) score because the new solution may have tweets having long-lengths and can convey important information.

E. Selection of Best $|P|$ Solutions for Next Generation

After generating the new population, $P^{'}$, it is merged with old population, P. Note that size of population $P^{'}$ equals to that of population P. Thus total solutions after merging will be $2 \times |P|$ out of which the best $|P|$ solutions in the objective space will be passed to the next generation. These best solutions are selected using the dominance and non-dominance relationships between the solutions in objective space. These relationships are calculated using the well known non-dominated sorting (NDS) and crowding distance based operators [18]. NDS algorithm assigns ranks to the solutions using their objective functional values and puts them in different fronts based on their rankings. Crowding distance operator determines which solution in a front lies in the more crowded region. For the selection of the best $|P|$ solutions, solutions are selected in a rank-wise manner until the number of solutions reaches the value $|P|$. In case of a tie, a solution having high crowding distance will be selected.

F. Updating SOM Training Data

In the next generation, SOM is trained using the newly generated solutions which have not seen before. It is important to note that the updated weight vectors of the neurons in the previous generation are now treated as initial weight vectors of the neurons in the next generation.
Example: If population P at t_{th} generation is $\{a, b, c\}$ and new population P' is $\{d, e, f\}$. Let after merging and selection of best solutions for the next generation $(t+1)$, new population P be $\{a, d, e\}$. Thus new SOM training data for $(t+1)$th generation will be $\{d, e\}$.

G. Termination Condition

The process of mating pool generation, crossover and mutation followed by selection and then updation of SOM training data is repeated until a maximum number of generations, g_{max} is reached. In the last generation, we will obtain a set of Pareto optimal solutions. A diamond box in Figure 1 shows this step.

H. Selection of Single Best Solution and Generation of Summary

At the end of the final generation, any MOO algorithm provides a set of non-dominated solutions (rank-1 solutions) on the final Pareto optimal front. All solutions in the final set have equal importance. Therefore, the decision maker has to select a solution based on his/her requirement. In this paper, two methods, supervised and unsupervised, are explored to select the best solution. Let us call these methods as $SBest$ and $UBest$, respectively.

- $SBest$: In this method, firstly, we will generate summaries corresponding to different solutions and then select that solution which has the highest ROUGE-1 score. Note that in calculation of ROUGE score, it makes use of gold/reference summary. However, in real time, the reference summary may not be available. That's why we have also explored unsupervised method. In this paper, using supervised method, our goal is to show that our proposed approach is able to generate a good summary from the dataset and by averaging results of all datasets corresponding to the best summary, we are able to beat the existing algorithms.
- $UBest$: In this method, an adaptive weighting scheme (AWS) [39] is utilized in which objective functional values are summed up after multiplying with their respective weights. The solution having the best value of the weighted sum will be considered as the best solution. Let $K \times \#Ob$ be the matrix of objective functional values, where, K and $\#Ob$ are the number of Pareto optimal solutions and number of objective functions used in our optimization strategy, respectively. Then, steps used to select the best solution are explained below:
  1) Normalize the values of objective functions by applying
  $$F_{kl} = \frac{Ob_{kl}}{Ob_l^+} \quad \text{where } Ob_l^+ = \max_{k \in K} Ob_{kl} \qquad (10)$$
  where, Ob_{kl} is the lth objective function value corresponding to kth solution.
  2) Construct the normalized weighted matrix by multiplying normalized objective function value with its respective weight as
  $$F\_wtd_{kl} = Ob_{kl} \times w_l \qquad (11)$$
  where, w_l is the weight factor assigned to lth objective.
  3) For each kth solution, evaluate the sum of weighted normalized objective functional values as defined below
  $$Score_k = \sum_{l=1}^{\#Ob} F\_wtd_{kl} \qquad (12)$$
  4) Find the solution having largest $Score$.

  Note that the weight factors can be determined after conducting a sensitivity analysis.

The tweets, in summary, are reported based on their occurrences in the original dataset. For example, the tweet which appears first in the dataset will be the first tweet in summary.

V. Experimental Setup

A. Datasets

In this paper, we have used the datasets related to the four disaster events, namely, (a) Sandy Hook elementary school shooting in USA (SH); (b) Uttarakhand's floods (UK); (c) Typhoon Hangupit in Philippines (TH), and (d) Bomb blasts in Hyderabad (HB). The number of tweets in these datasets are 2080, 2069, 1461, and, 1413, respectively. Same datasets are used in the paper [4]. Tweets in these datasets provide different relevant information like the number of casualties, and the current situation in various regions affected by the disaster, contact number of helping authorities and hospitals etc. The reference/gold summary is also available with these datasets which is utilized only for evaluation at the end of the execution of our proposed approach as our approach is fully unsupervised in nature. Calculation of objective functions is also fully unsupervised in nature. The other steps of the proposed approach do not consult any supervised information. Number of tweets in gold summary are 37, 34, 41 and 33 for SH, UK, TH, and HB datasets, respectively. Before passing any dataset as an input to our algorithm, some pre-processing steps are executed on the given datasets. These include removal of special characters, hash tag, stop words, user mentions and URL. Lower case conversion of all the words is also carried out.

B. Evaluation Measure

To check the performance/closeness of the generated summary with the actual summary, we have used ROUGE-N measure. It is measured using ROUGE Toolkit [40] (version 1.5.5). It counts the number of overlapping units between the generated summary with the actual summary. A summary having highest ROUGE score is considered more close to the actual summary. In our experiment, N takes the values of 1, 2, and L for ROUGE-1, ROUGE-2, and ROUGE-L, respectively. But, for comparison purpose with the existing algorithms, we make use of only ROUGE-2, and ROUGE-L scores as reference papers reported only these scores.

C. Parameters Used

Different parameter values used in our proposed framework are- *DE parameters*: $\mid P \mid = 25$, mating pool size=5, threshold probability in mating pool construction (β)=0.8, maximum number of generations (g_{max})=25, crossover probability (CR)=0.8, b=6, F=0.8. *SOM parameters*: initial learning rate (η_0)=0.6, training iteration in SOM=|P|, topology=rectangular 2D grid; grid size=$N_1 \times N_2 = 5 \times 5$, initial neighborhood size (σ_0)=$\frac{1}{2}\sqrt{(\sum_{i=1}^{m-1} N_i^2)/(m-1)}$. Sensitivity analysis on DE parameters and SOM parameters is discussed in the supplementary sheet available at https://github.com/nsaini1988/MicroblogSummarization/blob/master/Supplmentary.pdf. Minimum (S_{min}) and maximum (S_{max}) number of tweets to be in summary for SH, UK, TH and HB datasets are considered as $[34, 40]$, $[31, 37]$, $[39, 44]$, and $[31, 36]$, respectively. Word Mover Distance makes use of pre-trained word2vec model[1] to calculate the distance between two tweets. This model was trained on 53 million tweets related to various disaster events [41]. Results obtained are averaged over 5 runs of the algorithm. In most of the evolutionary based optimization algorithms [22], [42], generally, number of fitness function evaluations (NFE) is reported which is considered as a stopping criteria in these algorithms. It is equal to $|P| \times g_{max}$. In our case, NFE has a value of 625.

VI. Discussion of Results

In this section we will discuss the results obtained using supervised ($SBest$) and unsupervised ($UBest$) selection methods, comparison with exiting approaches and analysis of the results obtained.

A. Discussion of results obtained using $SBest$ selection method

In Table I, we have shown the average results over all datasets obtained by the proposed approach, MOOTweet-Summ, using both versions 'with SOM' and 'without SOM' based genetic operators. Various combinations of the objective functions (discussed in section III) are also explored to identify which set of objective functions is the best suited for our task. The corresponding results are reported in Table I. The best result was obtained by our approach when using 'without SOM' version with objective functions namely, maximum anti-redundancy (Ob_1) and tf-idf score (Ob_2). However, we have also reported different evaluation measures for each dataset in Table II. From this table, it can be analyzed that objectives, Ob_1 and Ob_2, are best suited for TH. While for SH and HB datasets, our approach attains good results when objectives Ob_1 and Ob_3 are used. As comparative approaches report the average results over four datasets, therefore, to make a fair comparison, we have reported the average results in Table IV in comparison with the state-of-the-art techniques.

In the literature, efficacy of SOM based reproduction operators (for constructing the mating pool) is already shown in solving various problems like automatic clustering [24], [43], [44], [45], document summarization [22], development of an evolutionary algorithm [25] etc. But, from the obtained experimental results, it is evident that effectiveness of the SOM based operators also depends on the datasets and problem statement chosen. SOM based operators are developed based on the assumptions that mating pool should be restricted to neighboring solutions of the current solution. This restricts the genetic operations to be performed between neighboring solutions only. Thus exploitation was preferred more over exploration. But in case of tweet-summarization, neighborhood of a neuron mostly consists of re-tweets. Thus if genetic operators are applied on re-tweets, then good quality solutions may not be generated. Thus, in this case, SOM based genetic operators only help in exploitation. But, our summarization task demands more exploration than exploitation. Therefore, our approach 'without SOM' based

[1] http://crisisnlp.qcri.org/lrec2016/lrec2016.html

TABLE I
AVERAGE ROUGE SCORES OVER ALL DATASETS ATTAINED BY THE PROPOSED METHOD USING SUPERVISED INFORMATION. HERE, † DENOTES THE BEST RESULTS; IT ALSO INDICATES THAT RESULTS ARE STATISTICALLY SIGNIFICANT AT 5% SIGNIFICANCE LEVEL.

Approach	SOM/Without SOM	Objective functions	Rouge-1	Rouge-2	Rouge-L
MOOTweetSumm	With SOM	Ob1+ Ob2+ Ob3	0.4912	0.2999	0.4850
		Ob1+ Ob2	0.4738	0.3033	0.4678
		Ob1+ Ob3	0.4843	0.3095	0.4790
	Without SOM	Ob1+ Ob2+ Ob3	0.4789	0.2984	0.4745
		Ob1+ Ob2	***0.4900***	***0.3150***	***0.4860***
		Ob1+ Ob3	0.4903	0.3192	0.4848

TABLE II
ROUGE SCORES OBTAINED BY THE PROPOSED APPROACH FOR DIFFERENT DATASETS USING $SBest$ SELECTION METHOD. BOLD ENTRIES INDICATE THE BEST RESULTS CONSIDERING 'WITH SOM' AND 'WITHOUT SOM' BASED OPERATORS.

		With SOM			Without SOM		
Event Name	Objective functions	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
SH	Ob1+ Ob2+ Ob3	0.5842	0.3612	0.5776	0.5940	0.3612	0.5874
	Ob1+ Ob2	0.5346	0.3303	0.5248	0.5842	0.3721	0.5842
	Ob1+Ob3	0.5842	0.3775	0.5743	**0.6139**	**0.3975**	**0.6073**
UK	Ob1+ Ob2+ Ob3	0.4400	0.2469	0.4329	0.4494	0.2577	0.4424
	Ob1+ Ob2	0.4423	0.2714	0.4376	**0.4541**	**0.2822**	**0.4447**
	Ob1+Ob3	0.4565	0.2791	0.4518	0.4471	0.2623	0.4400
TH	Ob1+ Ob2+ Ob3	**0.4181**	**0.2365**	**0.4097**	0.3697	0.2213	0.3655
	Ob1+ Ob2	0.3634	0.2158	0.3634	0.3845	0.2184	0.3782
	Ob1+Ob3	0.3866	0.2296	0.3802	0.3634	0.2241	0.3550
HB	Ob1+ Ob2+ Ob3	0.5223	0.3552	0.5198	0.5025	0.3534	0.5025
	Ob1+ Ob2	0.5198	0.3776	0.5173	0.5371	0.3914	0.5371
	Ob1+Ob3	0.5099	0.3517	0.5099	**0.5371**	**0.3931**	**0.5371**

TABLE III
ROUGE SCORES OBTAINED BY THE PROPOSED APPROACH FOR DIFFERENT DATASETS USING $UBest$ SELECTION METHOD. HERE, UNDER $UBest$, THE RESULTS OF SAW ARE SHOWN. BOLD ENTRIES INDICATE THE BEST RESULTS CONSIDERING 'WITH SOM' AND 'WITHOUT SOM' BASED OPERATORS.

	Ob1+Ob2+Ob3, With SOM			Ob1+Ob2, Without SOM		
Event Name	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
SH	0.5743	0.3848	0.5710	0.5842	0.3721	0.5842
UK	0.4376	0.2715	0.4329	0.4376	0.2577	0.4376
TH	0.3592	0.2019	0.3487	0.3592	0.1923	0.3487
HB	0.5223	0.3552	0.5198	0.3857	0.3914	0.5371
Average	0.4734	0.3033	0.4681	0.4417	0.3033	0.4769

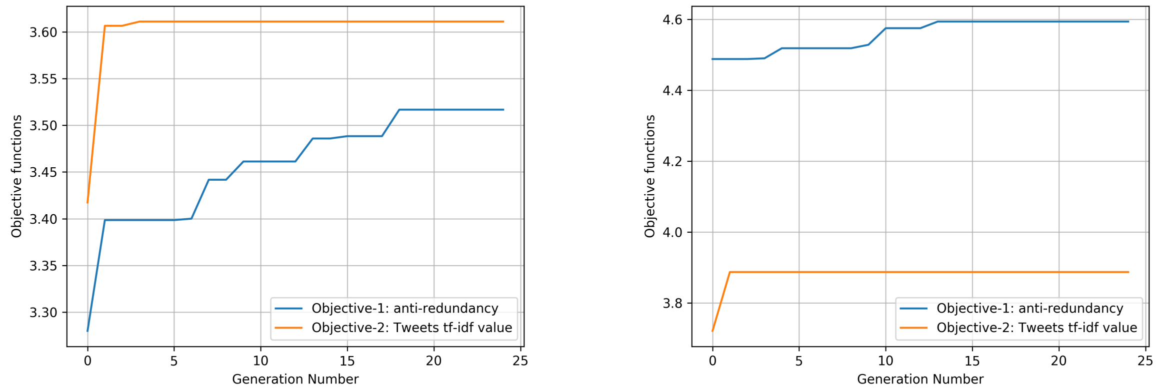
TABLE IV
AVERAGE ROUGE SCORES OVER ALL DATASETS ATTAINED BY EXISTING METHODS IN COMPARISON WITH THE BEST RESULTS OBTAINED BY THE PROPOSED APPROACH REPORTED IN TABLE I. HERE, WOSOM REFERS TO WITHOUT SOM, $SBest$ AND $UBest$ ARE THE SUPERVISED AND UNSUPERVISED SELECTION METHODS.

Approach	Rouge-2	Rouge-L
MOOTweetSumm (SBest, WOSOM, Ob1+Ob2)	**0.3150**†	**0.4860**†
MOOTweetSumm (SBest, SOM, Ob1+Ob2+ Ob3)	**0.2999**	**0.4850**
MOOTweetSumm (UBest, WOSOM, Ob1+ Ob2)	**0.3033**	**0.4769**
MOOTweetSumm (UBest, SOM, Ob1+ Ob2+ Ob3)	**0.3033**	**0.4681**
VecSim–ConComp–MaxDeg	0.1919	0.4457
VecSim–ConComp–MaxLen	0.1940	0.4506
VecSim–ConComp–maxSumTFIDF	0.1886	0.4600
VecSim–Community–maxSumTFIDF	0.1898	0.4591
ClusterRank (CR)	0.0859	0.2684
COWTS (CW)	0.1790	0.4454
FreqSum (FS)	0.1473	0.3602
Lex-Rank (LR)	0.0489	0.1525
LSA (LS)	0.1599	0.4234
LUHN (LH)	0.1650	0.4015
Mead (MD)	0.1172	0.3709
SumBasic (SB)	0.1012	0.3289
SumDSDR (SM)	0.0985	0.2602

genetic operators performs better than the 'with SOM' version.

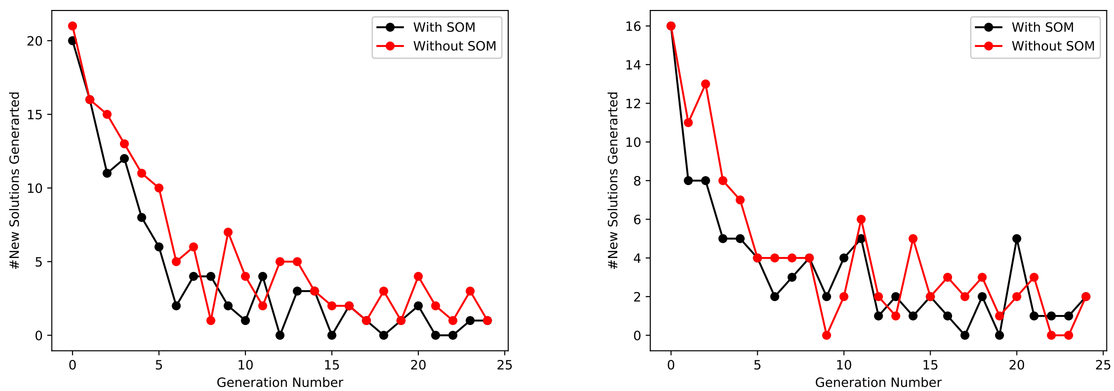
Exploration vs. Exploitation Behaviour: In Figs 3(a)-(b) for SH and HB datasets, respectively, exploration vs. exploitation behaviour of our proposed algorithm is shown with respect to the number of generations using two objectives, Ob_1+ Ob_2 (as it gives the average best result), with both versions, 'with SOM' and 'without SOM' based operators. Due to length restrictions, these behaviours are not shown for other two datasets. As can be seen by the red line corresponding to 'without SOM' based operator version, number of new solutions generated per generation is more as compared to 'with SOM' version, in most of the generations. That means 'without SOM' version explores the search space more efficiently. This is due to the random selection of three solutions out of the whole population to generate a new solution for the current solution as usually done in DE algorithm and thus, is able to provide the best average ROUGE score. However, both versions move towards exploitation as the number of new solutions generated is decreasing over the generations.

To check whether the used objective functions are optimized or not over the generations, we have plotted graphs showing generation wise maximum objective functional (Ob_1 and Ob_2) values for only two datasets, namely, SH and UK. These graphs are shown in Fig 2 which shows that objective



(a) (b)

Fig. 2. Generation wise objective function values using MOOTweetSumm (Without SOM, Ob1+ Ob2). Here, (a) and (b) correspond to SH and UK datasets, respectively.



(a) (b)

Fig. 3. Figures showing the number of new solutions generated over the generations by our proposed approach using two objectives, Ob_1+ Ob_2; a comparative study between 'with SOM' and 'without SOM' based operators. Here, (a) and (b) correspond to SH and HB datasets, respectively.

functional values are increasing over the iterations and become constant after a particular iteration due to limited length of the tweets and vocabulary size.

B. Discussion of results obtained using $UBest$ selection method

From the results shown in Table I, obtained using $SBest$ selection method, it can be analyzed that (i) in case of SOM-based operator, our approach performs better when all objectives functions are optimized simultaneously; (ii) in case of without SOM-based operator, our approach performs well when two objective functions, Ob1 and Ob2, are optimized simultaneously. Therefore, using the same set of objective functions for 'with SOM' and 'without SOM' based operators, we have explored the unsupervised method for selecting the best solution as discussed in section IV-H. The corresponding results are reported in the Table III. The weight factors assigned to different objective functions, Ob1, Ob2 and Ob3, when using SOM-based operator are 0.4, 0.3, and, 0.7, respectively. While in case of not using SOM-based operator, weight factors assigned to Ob1 and Ob2 are 0.3 and 0.7, respectively. Note that in case of SOM-based operator, weight values of 0.2, 0.3, and, 0.5 assigned to Ob1, Ob2 and Ob3, respectively, generate the same results. These weight factors are determined after conducting a thorough sensitivity analysis. On comparing the results of 'with SOM' and 'without SOM' based operators of $UBest$, both give ROUGE-2 score of 0.3033, but, in terms of ROUGE-L, proposed approach using 'without SOM' based operator is able to achieve 0.4769 which is higher than ROUGE-L score of 0.4681, obtained using with SOM-based operator. Note that ROUGE-L measures the matching of *longest common subsequence* between obtained summary and reference summary, thus, ROUGE-L can be more preferred than ROUGE-2. Similar discussions can be applied to ROUGE-1 score.

On comparing the best average ROUGE scores among $SBest$ and $Ubest$ selection methods, $SBest$ performs better than $Ubest$ which is quite obvious because of the use of supervised information. The best ROUGE-2 and ROUGE-L scores attained by $SBest$ are 0.3150 and 0.4860, respectively, while, using $UBest$, ROUGE-2 and ROUGE-L scores are 0.3033 and 0.4709, respectively. Thus, $UBest$ method is not able to reach to the exact results (average ROUGE score) obtained by $SBest$. But, it can be inferred that the results of $UBest$ are able to beat the results of existing approaches. Note that selection of a single best solution from the final Pareto optimal front is an active research area of MOO [39]. Researchers are exploring different techniques in this context.

C. Comparative Analysis

The comparative methods *VecSim–ConComp–MaxDeg, VecSim–ConComp–MaxLen, VecSim–ConComp–maxSumTFIDF, VecSim–Community–maxSumTFIDF* are based on ensembling technique, i.e., they consider the summary generated by different existing algorithms, and, then generate the final summary in an unsupervised/supervised way. Although this is a promising technique but very time-taking in the real-time scenario. Also, these approaches remove the redundant tweets before applying the ensembling algorithm. The remaining algorithms like Luhn, Lex-Rank, Mead etc. are very basic algorithms suggested in the literature [12]. The technique, COWTS, generates the summary based on the content words in the dataset. Our proposed approach is unique compared to all the existing approaches in the following ways:
1) Note that all the comparative methods used in the current paper do not provide the user a set of alternative solutions on the final Pareto front. Thus they do not provide the end-user an opportunity to select a single best summary out of many choices as per his/her requirement, while in our approach, there is a flexibility for the end-user to select a single one based on some objective functional value or his/her expert knowledge. 2) Moreover, unlike the other comparing approaches, redundant tweets are automatically removed from the resultant summary utilizing anti-redundancy objective function in our approach.

Experimental results suggest that our algorithm is able to beat all these algorithms as it attains the ROUGE-2 and ROUGE-L values of 0.3150 and 0.4860, respectively, using $SBest$ selection method. In other words, our algorithm improves by 62.37% and 5.65% in terms of ROUGE-2 and ROUGE$-$L scores, respectively, over the state-of-the-art techniques. Lex-Rank performs poorly among all techniques. Note that the *'improvement obtained'* is calculated using the formula ($\frac{ProposedMethod-OtherMethod}{OtherMethod} \times 100$).

D. Quality of Summaries for Different Solutions

To illustrate the qualities of summaries corresponding to different solutions on the final Pareto front obtained at the final generation using the proposed approach utilizing 'with SOM' and 'without SOM' based genetic operators, we have also plotted the ranges of Rouge-2/L score values attained by rank-1 solutions in Fig 6. We have chosen rank-1 solutions because the best solution belongs to this set. From Fig 6(a), (b), and, (d) for SH, UK, and, HB datasets, respectively, it can be analyzed that some solutions in 'without SOM' version have low ROUGE-2/L values but the best solution is identified by this version (as can be seen by Rouge values corresponding to green bullets). But, for UK dataset, median value of rank-1 solutions is high when using 'with SOM' version. Thus, it can be inferred that efficacy of SOM as a reproduction operator in summarization framework simply depends on the datasets used. Not in all cases, SOM based operators will be effective in solving the summarization task.

E. Pareto Fronts Obtained

Pareto fronts obtained by our proposed approach corresponding to the best results obtained using SBest selection method are shown in Fig 4 and 5 generated at the end of $\{0, 10, 20\}$th generation. The Pareto fronts obtained shown in Fig 4 and 5 are correspond to 'with SOM' and 'without SOM', respectively, for TH dataset. Due to limited space, we have not shown the Pareto fronts for other datasets. In the $0th$ generation, solutions are initialized randomly and thus randomly distributed over the objective space. On comparing with and without SOM version (Fig 4 and 5), it can be observed that using 'without SOM' version, we obtain more optimized and diverse set of solution which also support our results reported in Table I. In these figures, the '.' indicates a solution's objective functional values. Various colors represent different ranked or front solutions. Highest ranked solutions are indicated by color (blue) assigned to 'fr-0' as shown in the legend of Fig 4(a) and so on.

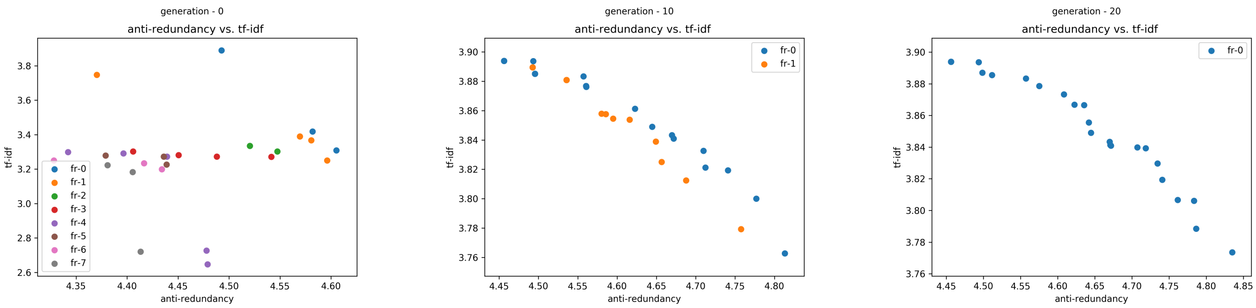
F. Analysis of Summarization Results Obtained

After careful manual inspection of gold summary and our predicted summary, following observations can be made:

1) Gold summary is prepared by human annotators. In general, redundant tweets should not be part of the summary. However, in actual summary (for example-gold summary corresponding to HB dataset), there exist $1-2$ redundant tweets. For example:
   - Lol at Indian Media showing picture of dead #MQM MPA Manzar Imam as terrorist suspect in Hyderabad blast.
   - RT @Ammar_Haider: Lol at Indian Media showing picture of dead #MQM MPA Manzar Imam as terrorist suspect in Hyderabad blast.

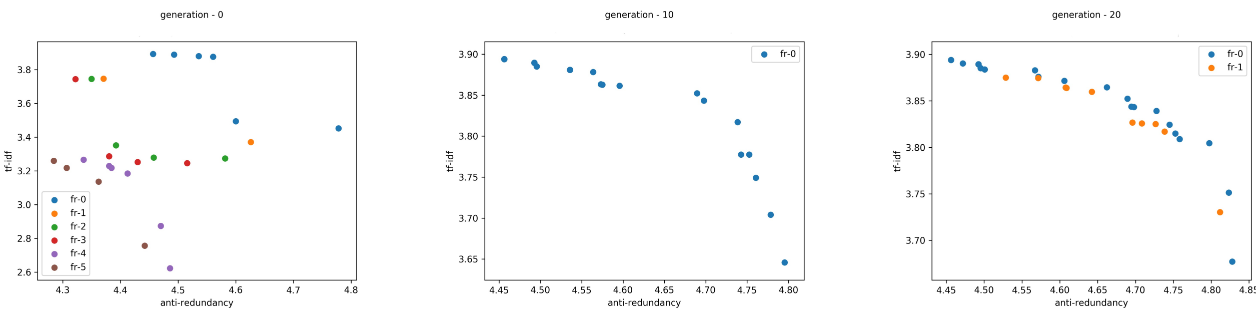
   Here the second tweet is the re-tweet of the first tweet.
2) Two tweets having high tf-idf score values (used as one of the objective functions) differ by only one word. But, only one is covered in the gold summary. While in predicted summary, both are covered. For example:
   - #Hyderabad blast proved once again d threat frm Muslim terrorist & Akhilesh Yadav needs to think twice before freeing *Muslims* lodged in jail.
   - #Hyderabad blast proved once again d threat frm terrorist & Akhilesh Yadav needs to think twice before freeing *terrorists* lodged in jail.

   Here, italic words are different. Further, to investigate this situation, we have computed the dis-similarities between these two tweets in semantic space. As these words have different word-vector representations, these two tweets have a dis-similarity (WMD) value of 0.14.



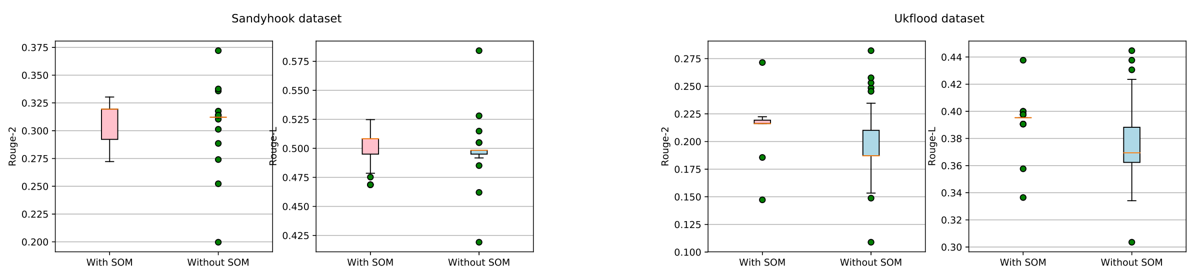
(a) (d) (c)

Fig. 4. Pareto optimal fronts obtained at the end of $\{0, 10, 20\}$th generation corresponding to TH dataset using 'With SOM' version.

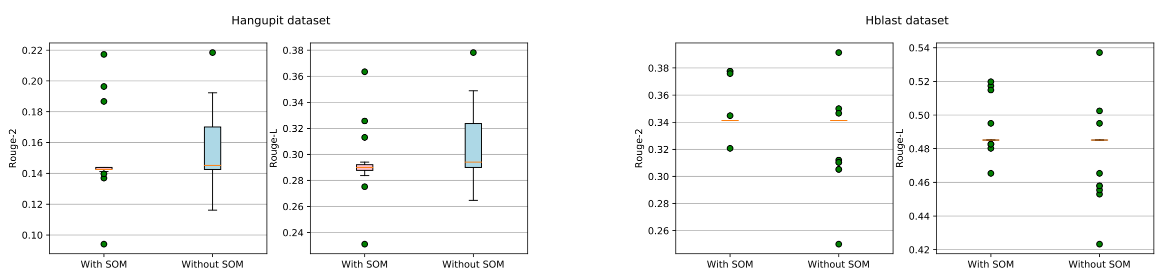


(a) (d) (c)

Fig. 5. Pareto fronts obtained at the end of $\{0, 10, 20\}$th generation corresponding to TH dataset using 'Without SOM' version.



(a) (b)



(c) (d)

Fig. 6. Box plots in sub-figures (a), (b), (c) and (d) for SH, UK, TH and HB datasets, respectively, illustrating the variations of average Rouge-2/Rouge-L values of highest ranked (rank-1) solutions of each document.

Moreover, they also have different tf-idf score values of 2.27 and 3.24, respectively. As these two tweets are dissimilar in semantic space and both of them have good tf-idf values (with respect to other tweets in the dataset), both are selected in the summary. Anti-redundancy objective does not help in this situation. To remove this drawback, in future, we would like to use some other sophisticated word embedding like BERT [46]. Moreover, we would also like to explore some tweet-specific embedding in association with emotion-aware embedding to better capture the semantic dissimilarities between tweets.

3) The best average result over all datasets (in Table IV) is obtained after optimizing anti-redundancy and maximum tf-idf score objectives. But, there exist some tweets having low tf-idf scores which are covered in actual summary (written by the human annotators) but not in the predicted summary. Therefore, there is a need to know the guidelines of generating gold summary which is not made available by the human annotators.

G. Statistical t-test

To validate the results obtained by the proposed approach, a statistical significance test named as, Welch's t-test [47], is conducted at 5% significance level. It is carried out to check whether the best average ROUGE scores (in Table IV) obtained by the proposed approach are statistically significant or occurred by chance. This t-test provides p-value. Minimum p-value signifies that our results are significant. The p-values obtained using Table IV are (a) $< .00001$ using ROUGE-2 score; (b) .000368 using ROUGE-L score. Test results support the hypothesis that obtained improvements by the proposed approach are not occurred by chance, i.e., improvements are statistically significant.

VII. An Application to Multi-document Summarization

To show the effectiveness of our proposed approach to other domain data, we have also performed multi-document summarization. The task is to generate fixed length summary (in terms of the number of words) given a collection of documents. For this task, we have used, DUC 2002, standard datasets provided by Document Understanding Conference. It contains 59 topics each having approx 10 documents. The corresponding multi-document summaries (two in number) each of 200 words are also available for each topic. Out of 59 topics, ten topics ranging from $d061j$ to $d070f$ from this dataset are considered while performing the experiments. The same set of topics are also considered in the comparative approaches (discussed below). The statistics about these specific topics (like number of lines etc.) are provided in the supplementary sheet.

A. Comparative Approaches and Differences with Our Approach

For the purpose of comparison, two existing evolutionary-based approaches are considered. The first approach utilized adaptive differential evolution [48] for optimization in which DE parameters are adaptive. In this approach, a weighted combination of two objectives, namely, anti-redundancy (AR)) and coverage (COV), is optimized. The mathematical definition of anti-redundancy is given in Eq. 4, while, coverage means the central theme of the document collection which should be covered in the summary. For a solution in the population, it is evaluated as $\sum_{i=1}^{N} sim(s_{vi}, \mathcal{O})$, where, N is the total number of sentences, s_{vi} is the vector representation (numeric vector) of the ith sentence belonging to the solution, $\mathcal{O}$ is the document vector calculated by averaging the sentence vectors. To represent the sentences in vector form, a well known tf-idf representation of vector space model in information retrieval [4], is utilized. To measure the similarity among sentences and sentences to document vector, cosine similarity is utilized.

In the second approach [49], these objectives (AR and COV) are optimized simultaneously (instead of using a weighted combination) and it uses well know genetic algorithm in the field of multi-objective optimization, i.e., non-dominating sorting genetic algorithm (NSGA-II) [18]. It also makes use of same sentence vector representation strategy and similarity measure as used in adaptive DE. But, in our approach, semantic similarity measure (WMD) is utilized. We do not make use of any vector representation scheme; therefore, in place of $\mathcal{O}$ in the coverage function definition, we have considered the representative sentence ($s_{\mathcal{R}}$) whose index in the document collection is evaluated by calculating the minimum average similarity of each sentence with other sentences in the topic, i.e.,

$$\arg\min_{\mathcal{R}} \left(\sum_{j=1, \mathcal{R} \neq j}^{N} dist_{wmd}(s_{\mathcal{R}}, s_j) \right) / (N-1) \qquad (13)$$

where, $\mathcal{R} = 1, 2 \ldots, N$, N is the total number of sentences, $dist_{wmd}$ is the word mover distance.

B. Results Obtained

In Table V, we have shown the results obtained for different topics in terms of ROUGE-2 measure. It can be observed that our proposed approach improves by 14.28% and 3.42% over adaptive DE (in short, ADE) and NSGA-II, respectively. ADE and our proposed approach both are based on differential evolution. ADE and NSGA-II use syntactic similarity, while, our approach uses semantic similarity. Note that WMD makes use of pre-trained *word2vec* model [37] on *Googlenews*[2] corpus which contains 3 billions words and each word vector is of 300 dimension. In the future, we want to see the effect of using vector representation of sentences in the semantic space.

TABLE V
Average ROUGE-2 scores corresponding to different topics of DUC2002.

Topic No. (↓)	Adaptive DE	NSGA-II	Proposed
$d061j$	0.266	0.306	0.337
$d062j$	0.188	0.200	0.200
$d063j$	0.245	0.275	0.220
$d064j$	0.194	0.233	0.392
$d065j$	0.144	0.182	0.183
$d066j$	0.201	0.181	0.258
$d067f$	0.239	0.260	0.286
$d068f$	0.491	0.496	0.294
$d069f$	0.184	0.232	0.220
$d070f$	0.224	0.262	0.332
Average	**0.238**	**0.263**	**0.272**

VIII. Conclusions and Future Work

In this paper, we have proposed a multi-objective based extractive summarization technique for solving the microblog

[2]https://github.com/mmihaltz/word2vec-GoogleNews-vectors

summarization task. A multi-objective binary differential evolution (MOBDE) technique is used as the underlying optimization strategy in the proposed summarization system. SOM-based operators are also explored in fusion with MOBDE. It utilizes the topological space identified by SOM to develop some new genetic (selection) operators. The similarity/dissimilarity between two tweets is calculated utilizing the word mover distance to capture the semantic information. Three objective functions are optimized simultaneously for selecting a good subset of tweets present in the dataset/event. Ablation study is also done to see which objective function combination performs the best for the given task. Results on 4 datasets related to disaster events prove the efficacy of the proposed technique compared to the state-of-the-art techniques in terms of better average ROUGE scores. Experimental results demonstrate that our proposed approach, MOOTweetSumm, has obtained 62.37% and 5.65% improvements over the existing techniques in terms of ROUGE-2 and ROUGE-L evaluations measures, respectively. Results are also validated using statistical significance test. The application of proposed approach is also shown for multi-document summarization task in which we have obtained 14.28% and 3.42% improvements over the two existing evolutionary-based techniques, ADE and NSGA-II, respectively.

In the future, we would like to investigate the effect of BM25 function (designed for short text) as a distance function between two tweets. We would also want to extend the current approach for online summarization of microblogging tweets.

ACKNOWLEDGMENT

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] I. Varga, M. Sano, K. Torisawa, C. Hashimoto, K. Ohtake, T. Kawai, J.-H. Oh, and S. De Saeger, "Aid is out there: Looking for help from tweets during a large scale disaster," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2013, pp. 1619–1629.

[2] G. Neubig, Y. Matsubayashi, M. Hagiwara, and K. Murakami, "Safety information mining—what can nlp do in a disaster—," in *Proceedings of 5th International Joint Conference on Natural Language Processing*, 2011, pp. 965–973.

[3] K. Rudra, S. Ghosh, N. Ganguly, P. Goyal, and S. Ghosh, "Extracting situational information from microblogs during disaster events: a classification-summarization approach," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 2015, pp. 583–592.

[4] S. Dutta, V. Chandra, K. Mehra, A. K. Das, T. Chakraborty, and S. Ghosh, "Ensemble algorithms for microblog summarization," *IEEE Intelligent Systems*, vol. 33, no. 3, pp. 4–14, 2018.

[5] N. Garg, B. Favre, K. Reidhammer, and D. Hakkani-Tür, "Clusterrank: a graph based method for meeting summarization," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[6] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of artificial intelligence research*, vol. 22, pp. 457–479, 2004.

[7] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2001, pp. 19–25.

[8] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of research and development*, vol. 2, no. 2, pp. 159–165, 1958.

[9] D. R. Radev, E. Hovy, and K. McKeown, "Introduction to the special issue on summarization," *Computational linguistics*, vol. 28, no. 4, pp. 399–408, 2002.

[10] A. Nenkova and L. Vanderwende, "The impact of frequency on summarization," *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005*, vol. 101, 2005.

[11] Z. He, C. Chen, J. Bu, C. Wang, L. Zhang, D. Cai, and X. He, "Document summarization based on data reconstruction." in *AAAI*, 2012.

[12] S. Dutta, V. Chandra, K. Mehra, S. Ghatak, A. K. Das, and S. Ghosh, "Summarizing microblogs during emergency events: A comparison of extractive summarization algorithms," in *Emerging Technologies in Data Mining and Information Security*. Springer, 2019, pp. 859–872.

[13] V. Gupta and G. S. Lehal, "A survey of text summarization extractive techniques," *Journal of emerging technologies in web intelligence*, vol. 2, no. 3, pp. 258–268, 2010.

[14] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, 2003, pp. 133–142.

[15] L. Wang, X. Fu, M. I. Menhas, and M. Fei, "A modified binary differential evolution algorithm," in *Life System Modeling and Intelligent Computing*. Springer, 2010, pp. 49–57.

[16] C. M. Fonseca and P. J. Fleming, "An overview of evolutionary algorithms in multiobjective optimization," *Evolutionary computation*, vol. 3, no. 1, pp. 1–16, 1995.

[17] S. Bandyopadhyay, S. Saha, U. Maulik, and K. Deb, "A simulated annealing-based multiobjective optimization algorithm: Amosa," *IEEE transactions on evolutionary computation*, vol. 12, no. 3, pp. 269–283, 2008.

[18] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *IEEE transactions on evolutionary computation*, vol. 6, no. 2, pp. 182–197, 2002.

[19] D. Zhang and B. Wei, "Comparison between differential evolution and particle swarm optimization algorithms." IEEE, 2014, pp. 239–244.

[20] R. Dong, "Differential evolution versus particle swarm optimization for pid controller design," in *Natural Computation, 2009. ICNC'09. Fifth International Conference on*, vol. 3. IEEE, 2009, pp. 236–240.

[21] J. Vesterstrom and R. Thomsen, "A comparative study of differential evolution, particle swarm optimization, and evolutionary algorithms on numerical benchmark problems." in *IEEE Congress on Evolutionary Computation*, vol. 2, 2004, pp. 1980–1987.

[22] N. Saini, S. Saha, A. Jangra, and P. Bhattacharyya, "Extractive single document summarization using multi-objective optimization: Exploring self-organized differential evolution, grey wolf optimizer and water cycle algorithm," *Knowledge-Based Systems*, vol. 164, pp. 45–67, 2019.

[23] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, no. 1, pp. 1–6, 1998.

[24] N. Saini, S. Chourasia, S. Saha, and P. Bhattacharyya, "A self organizing map based multi-objective framework for automatic evolution of clusters," in *International Conference on Neural Information Processing*. Springer, 2017, pp. 672–682.

[25] H. Zhang, A. Zhou, S. Song, Q. Zhang, X. Z. Gao, and J. Zhang, "A self-organizing multiobjective evolutionary algorithm," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 5, pp. 792–806, Oct 2016.

[26] H. Zhang, X. Zhang, X.-Z. Gao, and S. Song, "Self-organizing multiobjective optimization based on decomposition with neighborhood ensemble," *Neurocomputing*, vol. 173, pp. 1868–1884, 2016.

[27] R. Storn and K. Price, "Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces," *Journal of global optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[28] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *International Conference on Machine Learning*, 2015, pp. 957–966.

[29] M. A. H. Khan, D. Bollegala, G. Liu, and K. Sezaki, "Multi-tweet summarization of real-time events," in *Social Computing (SocialCom), 2013 International Conference on*. IEEE, 2013, pp. 128–133.

[30] L. Shou, Z. Wang, K. Chen, and G. Chen, "Sumblr: continuous summarization of evolving tweet streams," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2013, pp. 533–542.

[31] K. Rudra, N. Ganguly, P. Goyal, and S. Ghosh, "Extracting and summarizing situational information from the twitter social media during disasters," *ACM Transactions on the Web (TWEB)*, vol. 12, no. 3, p. 17, 2018.

[32] C. De Maio, G. Fenza, V. Loia, and M. Parente, "Time aware knowledge extraction for microblog summarization on twitter," *Information Fusion*, vol. 28, pp. 60–74, 2016.

[33] K. Rudra, P. Goyal, N. Ganguly, P. Mitra, and M. Imran, "Identifying sub-events and summarizing disaster-related information from microblogs," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 2018, pp. 265–274.

[34] K. Rudra, A. Sharma, N. Ganguly, and S. Ghosh, "Characterizing and countering communal microblogs during disaster events," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 2, pp. 403–417, 2018.

[35] G. Liang, W. He, C. Xu, L. Chen, and J. Zeng, "Rumor identification in microblogging systems based on users' behavior," *IEEE Transactions on Computational Social Systems*, vol. 2, no. 3, pp. 99–108, 2015.

[36] S. Madisetty and M. S. Desarkar, "A neural network-based ensemble approach for spam detection in twitter," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 4, pp. 973–984, 2018.

[37] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[38] S. S. Haykin, S. S. Haykin, S. S. Haykin, and S. S. Haykin, *Neural networks and learning machines*. Pearson Upper Saddle River, NJ, USA:, 2009, vol. 3.

[39] Z. Wang and G. P. Rangaiah, "Application and analysis of methods for selecting an optimal solution from the pareto-optimal front obtained by multiobjective optimization," *Industrial & Engineering Chemistry Research*, vol. 56, no. 2, pp. 560–574, 2017.

[40] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Text Summarization Branches Out*, 2004.

[41] M. Imran, P. Mitra, and C. Castillo, "Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA), may 2016.

[42] S. Khalilpourazari and S. H. R. Pasandideh, "Multi-objective optimization of multi-item eoq model with partial backordering and defective batches and stochastic constraints using mowca and mogwo," *Operational Research*, pp. 1–33, 2018.

[43] N. Saini, S. Saha, and P. Bhattacharyya, "Automatic scientific document clustering using self-organized multi-objective differential evolution," *Cognitive Computation*, Dec 2018. [Online]. Available: https://doi.org/10.1007/s12559-018-9611-8

[44] N. Saini, S. Saha, A. Harsh, and P. Bhattacharyya, "Sophisticated som based genetic operators in multi-objective clustering framework," *Applied Intelligence*, vol. 49, no. 5, pp. 1803–1822, 2019.

[45] N. Saini, S. Saha, C. Soni, and P. Bhattacharyya, "Automatic evolution of bi-clusters from microarray data using self-organized multi-objective evolutionary algorithm," *Applied Intelligence*, 2019 (accepted).

[46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[47] B. L. Welch, "The generalization ofstudent's' problem when several different population variances are involved," *Biometrika*, vol. 34, no. 1/2, pp. 28–35, 1947.

[48] R. M. Alguliev, R. M. Aliguliyev, and C. A. Mehdiyev, "Sentence selection for generic document summarization using an adaptive differential evolution algorithm," *Swarm and Evolutionary Computation*, vol. 1, no. 4, pp. 213–222, 2011.

[49] H. H. Saleh, N. J. Kadhim, and A. A. Bara'a, "A genetic based optimization model for extractive multi-document text summarization," *Iraqi Journal of Science*, vol. 56, no. 2B, pp. 1489–1498, 2015.

Naveen Saini He is currently working towards the Ph.D. degree at the Indian Institute of Technology Patna, Bihar, India. He received the M.Tech. degree in Information Technology from the YMCA University of Science and Technology, Faridabad, Haryana in 2014. His research interests include multi-objective optimization, pattern recognition, clustering and natural language processing. His area of research broadly covers many facets of extractive summarization. He has published research papers in Cognitive Computation, Knowledge-based systems, ACM Transactions on Multimedia Computing Communications and Applications, PLOS ONE, Applied Intelligence and in some peer reviewed international conferences.

Sriparna Saha She is currently an Associate Professor in the Department of Computer Science and Engineering, Indian Institute of Technology Patna, India. Her current research interests include machine learning, pattern recognition, multi-objective optimization, language processing and biomedical information extraction. She received her masters' and Ph.D. degrees in Computer Science from Indian Statistical Institute Kolkata, India, in the years 2005 and 2011, respectively. She has authored or coauthored more than 120 papers.

She is the recipient of Lt Rashi Roy Memorial Gold Medal from the Indian Statistical Institute for outstanding performance in M.Tech. (computer science), Google India Women in Engineering Award, 2008, NASI Young Scientist Platinum Jubilee Award 2016, BIRD Award 2016, IEI Young Engineer's Award 2016, SERB Women in Excellence Award 2018 and SERB Early Career Research Award 2018 and many others. She is the recipient of Humboldt Research Fellowship, Indo-U.S. Fellowship for Women in STEMM (WISTEMM) Women Overseas Fellowship program 2018 and CNRS fellowship. She had also received India4EU fellowship of the European Union to work as a Post-doctoral Research Fellow in the University of Trento, Italy from September 2010-January 2011. She is also the recipient of Erasmus Mundus Mobility with Asia (EMMA) fellowship of the European Union to work as a Post-doctoral Research Fellow in the Heidelberg University, Germany from September 2009-June 2010.

Pushpak Bhattacharyya He is a computer scientist and a Professor at Computer Science and Engineering Department, IIT Bombay. Currently, he is also the Director of Indian Institute of Technology Patna. He is a past President of Association for Computational Linguistics (2016-17), and Ex-Vijay and Sita Vashee Chair Professor. He currently heads the Natural language processing research group "Center For Indian Language Technology (CFILT)" lab at IIT Bombay.

He completed his undergraduate studies (B. Tech.) from IIT Kharagpur and Masters (M.Tech) from IIT Kanpur. He finished his Ph.D. from IIT Bombay in 1994. His research areas are Natural Language Processing, Artificial Intelligence, Machine Learning, Psycholinguistics, Eye Tracking, Information Retrieval, and Indian Language WordNets - IndoWordNet. A significant contribution of his research is Multilingual Lexical Knowledge Bases like IndoWordNet and Projection. He is the author of the textbook 'Machine Translation'. He has led government and industry projects of international and national importance and has received faculty grants from IBM, Microsoft, Yahoo, and the United Nations. He has authored or coauthored more than 100 papers in top-tier NLP conferences like ACL, COLING, NAACL, CoNLL, EMNLP, LREC etc. He is a Fellow of Indian National Academy of Engineering (FNAE). For sustained contribution to technology, he received the Manthan Award of Ministry of IT (2009), P.K. Patwardhan Award of IIT Bombay (2008) and VNMM Award of IIT Roorkey(2014).