The Journal of Supercomputing (2023) 79:5013–5036
https://doi.org/10.1007/s11227-022-04842-4

Opinion texts summarization based on texts concepts with multi-objective pruning approach

Sajjad Jahanbakhsh Gudakahriz[1] · Amir Masoud Eftekhari Moghadam[1] · Fariborz Mahmoudi[2]

Accepted: 20 September 2022 / Published online: 7 October 2022

Abstract

Considering the huge volume of opinion texts published on various social networks, it is extremely difficult to peruse and use these texts. The automatic creation of summaries can be a significant help for the users of such texts. The current paper employs manifold learning to mitigate the challenges of the complexity and high dimensionality of opinion texts and the K-Means algorithm for clustering. Furthermore, summarization based on the concepts of the texts can improve the performance of the summarization system. The proposed method is unsupervised extractive, and summarization is performed based on the concepts of the texts using the multi-objective pruning approach. The main parameters utilized to perform multi-objective pruning include relevancy, redundancy, and coverage. The simulation results show that the proposed method outperformed the MOOTweetSumm method while providing an improvement of 11% in terms of the ROGUE-1 measure and an improvement of 9% in terms of the ROGUE-L measure.

Keywords Opinion texts summarization · Concepts of texts · Multi-objective pruning · Relevancy · Coverage · Redundancy

1 Introduction

Communications on social media result in creating and sharing a large volume of data in various formats, including text, audio, image, and video. This large volume of data can be utilized to extract the patterns and behaviors in these media platforms. The main data generated and shared on these platforms is in text format. This

✉ Amir Masoud Eftekhari Moghadam
eftekhari.moghadam@gmail.com

[1] Department of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran

[2] Advanced Analytics Department, General Motors Company, Warren, MI, USA

data needs to be summarized and downsized for easier understanding, thus making it easily accessible for various applications [1]. There are a large number of applications, and the possibility of analyzing the opinions expressed inside the sheer volume of texts on social media will make these applications beneficial. As a result, it is desired to be able to use automatic systems to create a summary of opinion texts published on social networks and provide users with these summaries [2, 3].

Opinions and sentiments are often expressed by people with relevant experiences. A customer may seek the opinions of others before purchasing an item or deciding to watch a specific movie to gain an attitude about that specific action based on the experiences of others. Further, people involved in businesses can make accurate and optimized decisions once they understand the opinions of the users and/or customers [4]. Considering the large volume of opinion texts published on social networks, it is very difficult for individuals to easily evaluate and utilize these opinions. Opinion mining, also known as sentiment analysis, is one of the most active research fields in the field of natural language processing and computer science in the last decade. The purpose of opinion is to define automatic tools that can extract sentimental information from opinion texts. Opinion mining or sentiment analysis has different areas, such as sentiment classification [5], feature extraction [6], and opinion text summarization [7]. Summarization of opinion texts can significantly help in benefiting from these sentiments. In essence, after producing a summary of opinion texts, users can easily and quickly make use of these texts. As such, there is growing interest among researchers to develop new methods for the summarization of opinion text [8, 9].

In general, summarization techniques can be classified into two main categories based on their approach, i.e., syntax-based and semantic-based [10, 11]. The first approach makes use of a syntactic parser to analyze and represent the text according to the grammar. In contrast, the main goal of the semantic-based approach is to perform summarization based on the semantic representation of the text. In the syntax-based approach, the syntactic structure of the sentences, the text, or text segments are identified after which summarization is performed based on the identified structure. To determine the syntactic structure, methods such as parse trees and graphs are used. In the semantic-based approach, the semantic meaning of the sentences, the text, or text segments are identified, which forms the basis of the summarization. The main limitation of the syntax-based approach is the lack of semantic representation of the initial text. On the other hand, the main limitation of the semantic-based approach is its dependence on human expertise to create an anthology of the domain and the rules.

As noted earlier, the goal in the summarization of opinion texts is to receive a set of opinions expressed on a social network to create a useful summary that includes the content of the majority of initial texts [1, 12, 13]. Summarizing the opinions and sentiments expressed on social networks is a cutting-edge research area where a large number of studies have been performed to enhance linguistic quality and reduce the redundancy of summarization methods. However, the summarization of opinion texts faces numerous challenges due to the complexities of natural language processing as well as the high complexity and large volume of the texts [8]. As a result, it is necessary to develop a method that can perform summarization by reducing the complexity of the texts and can provide a summary with acceptable quality.

Hence, the main objective of the current study is to propose a method for summarizing opinion texts that can provide acceptable summarization accuracy and quality compared to available methods.

The proposed method has two main contributions:

- We present a new method for the summarization of opinion texts based on the concepts of the texts and dimension reduction in the clustering step.
- In the proposed approach, summarization is performed using the multi-objective pruning mechanism based on relevancy, redundancy, and coverage parameters.

The rest of the paper is organized as follows: Sect. 2 reviews the relevant literature and background on the field of this paper. Section 3 describes the proposed method for the summarization of opinion texts in detail. Section 4 presents the simulations performed to evaluate the efficiency of the proposed method and the results of these simulations. Finally, we conclude the paper in Sect. 5 and the extension of the proposed method is also illustrated.

2 Literature review and taxonomy of text summarization

This section reviews the relevant literature and the background in the field of this paper. As the first step, various text summarization methods and the taxonomy of these methods are discussed. Then, previous studies focusing on the subject are briefly reviewed.

2.1 Text summarization taxonomy

Automatic text summarization has gained the attention of researchers for several decades. A text summary consists of one or more texts and covers the important information of the initial text or texts. However, the length of the summary is less than half of the length of the original text or texts, and it usually has a much smaller size [14]. Different studies vary in their approach toward summarization. Some create the summary in a visualized or statistical manner, while some consider text summaries [8]. Summarization methods are classified based on different criteria, as shown in Fig. 1.

One of the types of text summarization involves single-document or multi-document summarization [15]. In single-document summarization, the summary is created from the content of a single text, while multi-document summarization uses various texts to produce the summary [8]. In terms of the used languages, there are three types of summaries, i.e., mono-lingual, multi-lingual, and cross-lingual [16]. When the language of the text being summarized and the language of the summarized text are identical, the summarization would be considered mono-lingual. However, when the text being summarized includes several languages and summarization is performed in one of these languages, the summarization would be considered multi-lingual. Finally, when the text being summarized is in one language

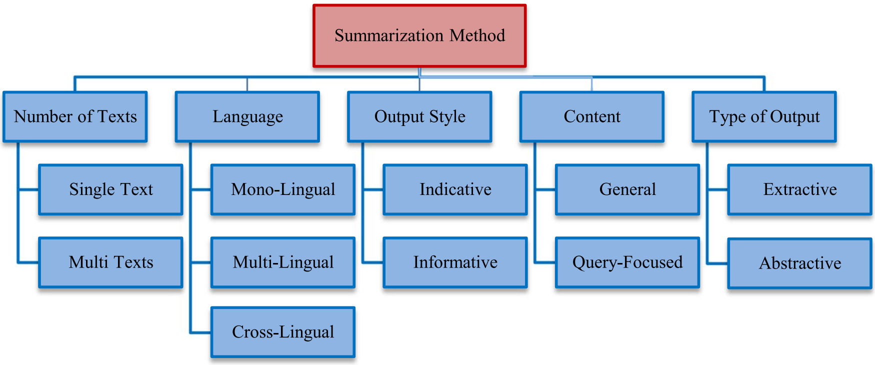

Fig. 1 Types of summarization methods

and the summarization is performed in another language, the summarization is considered cross-lingual. Based on the output style, there are indicative and informative summaries as well [17]. Indicative summaries are extracted in a way that they can express the subject matter of the text. This is while informative summaries are extracted in such a way that they can express the content and the information contained in the text. In addition, the produced summary can also be divided based on content into two main categories, i.e., general or query-focused (or topic-focused) and user-focused [18]. In query-focused summaries, the summary is produced based on queries (or questions) related to the content, while in general summaries, the summary is produced based on general perceptions of information.

Another important and useful type of text summarization is based on the type of output. In this type, there are two types of summarization, i.e., extractive summarization and abstractive summarization [19]. In extractive summarization, representative sentences are extracted from the text to express the main content of the text. The importance of dependent sentences will depend on the statistical and linguistic characteristics of those sentences [10]. In abstractive summarization, a summary of the text is created that includes words and sentences different from those inside the original text; however, the content of the produced summary will express the content of the original text. In essence, in this type of summarization, the summary is produced using new sentences that express the content of the original text [20].

2.2 Previous works

This section reviews several previous works focusing on text summarization. Some of the studies deal with summarization as a single-objective problem. These works include [21–29], where a static attribute is used for assigning value to the texts. In [21], Rudra et al. determine the value of the texts based on the coverage of important content words, such as nouns, verbs, and numerals. In [22], Dutta et al. proposed a mixed model combining several basic summarization algorithms to provide a summary better than the one produced by each basic algorithm. In [23], Garg et al. present a clustering-based method for summarization while using a centroid-based approach for assigning value to the texts inside the clusters. In [24], Erkan and Radev proposed a graph-based stochastic method. An intra-sentence cosine similarity criterion was used as the weight of the edges in the graph representation of the sentences, while a similarity matrix was created using the similarity criterion. At last, a thresholding mechanism was applied to identify and extract the sentences with the highest importance from the similarity matrix. In [25], Gong and Liu proposed an unsupervised method to extract information, including collocations and shared words used in different sentences. The input texts are converted into a matrix, where the rows indicate single words and columns indicate a sentence. Finally, SVD [30] is applied to this matrix to produce the summary. In [26], Luhn identifies descriptive words by determining high-frequency and low-frequency thresholds. Accordingly, words with a frequency higher than the high-frequency and lower than the low-frequency are eliminated, while the rest of the words are selected as descriptive words that express the important content of the original text. In [27], Radev et al. presented a centroid-based multi-document summarization method. As the first step, the topics are identified using agglomerative clustering. Then, a centroid-based method is used for identifying the important words in each cluster. In [28], Nenkova and Vanderwende proposed a multi-document frequency-based summarization method. In this method, each sentence is assigned a score based on the average likelihood of the presence of words in the sentence, followed by selecting the sentences with the best scores. In [29], Zhanying et al. presented a framework for summarization based on data reconstruction, where sentences that can best reconstruct the entire original text are selected as the summary.

Moreover, in some studies, summarization is dealt with as a multi-objective problem. Some of these studies include [31–33]. In [31], Algholio et al. present a single-document extractive summarization method. At first, the sentences in the original text are clustered using the K-Means algorithm to discover all the topics present in the text. Afterward, to select the important sentences in the clusters, an optimization model is presented that optimizes an objective function that uses means and the harmonics of objective functions that meet the coverage and diversity of the sentences selected for the summary. In [32], Chakrabouti et al. considered tweet summarization as a multi-objective optimization problem. Three key characteristics, i.e., relevance, diversity, and coverage, were selected as the objectives of the optimization, where the goal is to optimize these characteristics in addition to summarization. In [33], Siney et al. presented a method based on multi-objective optimization for tweet summarization. Various criteria, including length, TF-IDF, lack of redundancy, and the measurement of different aspects of the summary, were simultaneously optimized using the querying capability of a multi-objective differential evolution technique.

Works carried out in [34–39] focus on clustering for summarization. In [34], Dutta et al. presented an extractive method for tweet summarization based on community detection on the similarity graph of the tweets. In [35], Zhu et al. proposed an opinion-mining system on Chinese microblogs called the CMiner. After obtaining the aspects of the opinion, the goals of the opinion are clustered into a number of groups, the representative goals are extracted, and summarization is performed for each group. In [36], Jabrakumar et al. proposed an extractive method for the summarization of short texts published on microblogs using the clustering technique. To identify prioritized and important texts in each cluster, the closed words model is used and the presence of smiley faces, hashtags, and emphatic words is also considered. To mitigate the challenge of summarizing short texts in [37], Neu et al. proposed a new method that uses BM25 to assign weights to each short text and syntactic parsing to produce important information. In [38], Waheeb et al. present an unsupervised method for the summarization of multi-document Arabic texts, where clustering and the Word2Vec model were used for reducing redundancy. To represent and store texts based on meaning, the Word2Vec model is used, followed by using the K-Means algorithm and a cosine similarity criterion to select distinct documents from each set based on the distance criterion. In [39], Yusho presented a method for opinion summarization based on topic clustering. At first, the clustering of the topic is performed on noun-adverb clauses based on the model. Then, a number of noun-adverb clauses are selected from each cluster to create the summary.

The semantic-based approach is one of the important methods for sentiment summarization. The work presented by Labourt et al. in [40] is among the few studies focusing on the conceptual approach as an attempt to perform abstractive summarization of opinions. This method simplifies the syntax of the sentences, reproduces the sentences, and provides conceptual representations of the sentences to complete the summarization process. In [41], Amplayour and Sonagh presented a new method for the summarization of opinion texts based on the model. In the proposed method, a model is created for classifying sentiments and another model for extracting the aspects of the opinion. Then, by combining the outputs of these two models, the summarization of the opinions is performed. In [42], Bahatia et al. presented a method for query-focused extractive summarization based on aspects using PCA. In the proposed method, the main aspects are first extracted using dependency rules. Then, the opinion related to each aspect is extracted from each sentence. In [43], Raul and Maho proposed a hierarchical summarization method for summarizing large opinion texts in a couple of sentences. The proposed method first summarizes the opinion texts into one sentence using four basic methods, including SumBasic, LSA, TextRank, and LexRank. Moreover, machine learning algorithms were also considered in the summarization process. In [44], Abdi et al. presented a method for the summarization of opinion texts. This method is an extractive method that utilizes machine learning. The presented method makes use of prior knowledge to identify the class and intensity of each sentiment. Then, the important characteristics present in the sentiments are extracted, and this information is used for extracting the important sentiments. In [45], Mauli worked on summarizing long opinion texts. The proposed method summarizes a long emotional text in such a way that the main expressed sentiments are maintained and the readability of the text is not reduced.

3 The proposed method

The main problem in this paper is to produce a summary of the opinion texts published on a website, social network, or any other social media. However, the complex nature of the opinion texts published on social media makes their summarization

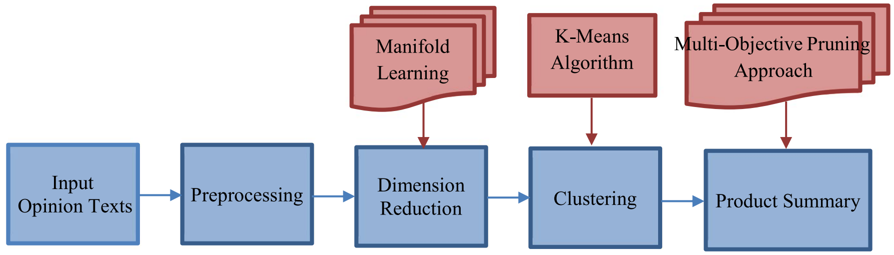

Fig. 2 General view of the proposed approach

a highly challenging task. The generated summary must be of high quality, must express the content of the majority of the topics present in the text, and must include the minimum amount of redundancy. Thus, the proposed approach has the following characteristics:

- Use of an appropriate process for reducing the complexity of opinion texts
- The possibility of representing opinion texts based on their intrinsic dimensions
- The possibility of producing a summary based on the concepts of the text
- The produced summary includes the highest levels of coverage and relevance as well as the lowest level of redundancy

To process and summarize opinion texts, these texts should be represented by a natural language representation model. This representation creates high dimensions [46], which negatively impacts the efficiency of opinion mining algorithms and techniques. Meanwhile, short texts quickly spread through various websites and social networks, resulting in a great volume of text that must be efficiently analyzed [47]. The next challenge facing opinion text summarization is that the generated summaries often suffer from high levels of redundancy [48]. Repeated and overlapping texts that cover distinct concepts result in the high redundancy of the produced summary. In the proposed method, dimension reduction is used for mitigating the challenge caused by the complexity of the texts, while the multi-objective pruning approach is employed for producing high-quality summaries. The general view of the proposed method is depicted in Fig. 2.

To describe the steps of the proposed method, it is necessary to state that the proposed method in this paper is one of the methods based on clustering. In these methods, after acquiring the texts and preprocessing, clustering is done on the texts, then summarization is done on the clusters. To produce better clusters, in this paper, dimension reduction is used before clustering. Also, in the summary production step, summarization is done based on the concepts of the texts and with a multi-objective pruning approach. The proposed method uses manifold learning to reduce the dimensions of the opinion texts in an attempt to identify the intrinsic dimensions of the texts instead of dealing with texts with high dimensions. After reducing the dimensions and identifying the intrinsic dimensions of the texts, clustering will be performed. From the texts in each cluster, the text or texts that best represent the remaining texts will be selected and added to the extracted summary. In the method proposed in this paper, the multi-objective pruning approach is used based on the concepts of the texts. To select the texts from the clusters, instead of selecting important and valuable texts, an iterative process is employed for multi-objective pruning of low-importance texts, texts with low levels of informational loading, and texts whose information is present inside other texts. Afterward, the texts remaining in each cluster will be selected. Finally, multi-objective pruning is performed on the summaries of the clusters, resulting in the final summary.

3.1 The details of the proposed method

The proposed method in this paper receives several mono-lingual opinion texts and selects a number of texts as summaries from these texts. The selected texts should contain the information of the majority of the input texts. Therefore, the method presented in this paper is multi-document in terms of the number of texts, mono-lingual in terms of language, informative in terms of output style, general in terms of content, and extractive in terms of the output type. Extractive summarization has several benefits, including the independence of the domain. Further, the summary produced in this method has high informative [48]. In order to perform extractive summarization, similar texts must be identified, and a number of texts that express the content of the majority of the texts must be selected based on the similarity among the texts. The method proposed here for summarizing opinion texts is an extractive method that uses the manifold learning algorithm to reduce the complexity of opinion texts and multi-objective pruning of texts based on the text's concepts. Figure 3 depicts the proposed method in detail.

3.1.1 Acquiring opinion texts and preprocessing

As the first step, the opinion texts to be summarized are acquired. The datasets used in this study include tweets collected from Twitter where each dataset includes the texts related to a single event. In order to perform any type of analysis on text data, it must be preprocessed [49]. Accordingly, the next step after acquiring the texts is to perform preprocessing on the text to mitigate the preliminary problems that may be present in the texts as much as possible. In fact, in the preprocessing step, unnecessary and noisy items are removed from the texts and the texts are converted into a structured form as much as possible so that the processing and analysis of the texts can be done with high quality and more easily. The preprocessing performed in this study has two phases. The first phase eliminates unwanted and noisy elements from the texts which involves the following: removing repeated letters, emojis, emails, Twitter signs, hashtags, and extra spaces. Since opinion texts are often written informally and ungrammatically, they have a weak structure, so it is necessary to convert them into a structured format, which is done in the second phase of preprocessing.

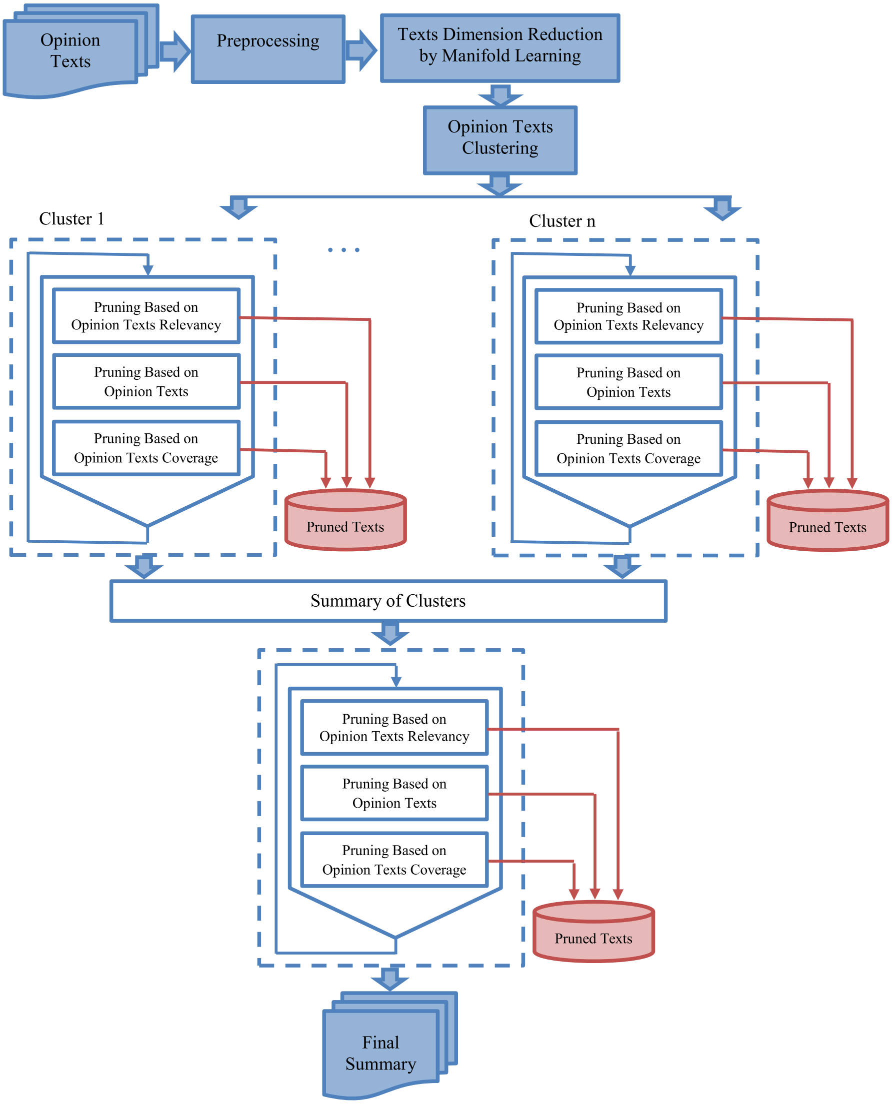

Fig. 3 Details of the proposed method

The second phase utilizes methods to convert unstructured texts into structured ones. These methods include acronym conversion, lowercasing, and lemmatizing.

3.1.2 Dimension reduction and clustering of opinion texts

After preprocessing, the texts must be represented using a model. Considering the wide range of words and the large size of the word dictionary, the utilized model will have high dimensionality. Thus, the vector created for the attributes will be very

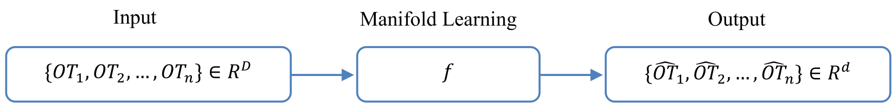

Fig. 4 The model of opinion texts dimension reduction by manifold learning

long and sparse. These high dimensions and the sparsity of the vector make the processing performed on them have a high computational load, preventing their easy understanding. In order to perform appropriate and highly efficient processing on the created vectors, they go through a dimension reduction process to identify their intrinsic dimension.

Manifold learning has attracted extensive attention in various fields for dimension reduction and data representation; however, it has rarely been used for text summarization [50]. Using this method, the intrinsic parameters that are the main factors for distinguishing data can be identified and the interrelationships among the data can be expressed in a space with lower dimensions. Using manifold learning to identify the intrinsic dimensions, the texts can be modeled as displayed in Fig. 4.

According to the above figure, the primary opinion texts are displayed in the D dimensions. Manifold learning maps this D-dimensional data to d-dimensional space (d < < D) in such a way that the intrinsic nature of the opinion texts is preserved and the distance of the dimension reduced texts is also preserved. In this paper, opinion texts are converted into vector form by Doc2Vec model, then these texts are reduced in dimension by ISOMP algorithm, which is one of the common algorithms of manifold learning, and the texts are converted into vector form with fewer dimensions.

After reducing the dimensions of the opinion texts using manifold learning, clustering is performed on these texts. Clustering of opinion texts causes texts with high similarity in terms of semantics and sentiment to be placed in a cluster, then summarization is done based on the clusters. The K-Means algorithm is the most common clustering algorithm due to its simplicity and efficiency, but it doesn't perform well when faced with high-dimensional data due to the presence of noise and redundant features. For this reason, in the proposed method after dimension reduction, K-Means algorithm is used for clustering, and groups the opinion texts into a number of clusters. In this paper, the number of clusters is determined based on the dataset and the value of K using the bisecting method.

3.1.3 Multi-objective pruning based on the texts concepts

Since clustering is a common task in most summarization methods, after clustering and identifying similar texts, we can use two approaches to produce summaries from each cluster. In the first approach, the syntax structure of the texts is considered and summarization is performed based on the roles of the words in the texts. In the second approach, the semantic structure of the texts is considered and summarization is performed. Previous methods have often focused on the syntactic structure of the texts, not paying much attention to the semantic structure for producing summaries. Furthermore, the majority of previous methods utilize the position of the sentences inside the texts along with probability formulae to calculate the value of the sentences. In the proposed method, the semantic structure of the texts is considered and summarization is performed based on concepts. In essence, the concepts of texts in each cluster are considered to perform summarization on each cluster.

This paper uses the YAKE algorithm [51] to extract keywords to be used as the main concepts of the texts. YAKE is a light-weight unsupervised automatic keyword extraction method that rests on text statistical features extracted from single documents to select the most important keywords of a text. Firstly, this algorithm utilizes local statistical characteristics to determine the importance of single-word phrases extracted from the text. Then, a specific n-gram model is applied to create multi-word phrases, using an exploratory measurement process to determine their relationships.

After identifying the concepts in each text, the texts in each cluster can be summarized based on the concepts present in those texts. The summarization will be performed using the multi-objective pruning approach. In the pruning approach, instead of a single step of selecting the texts, the texts can be gradually pruned and the final texts can be obtained in a step-by-step process. The approach used for pruning in the current study is a multi-objective approach. In the multi-objective approach, pruning is performed based on different objectives and parameters to be able to evaluate the quality of the produced summary from different aspects. After extracting the concepts of the texts in the clusters, multi-objective pruning is performed based on the following parameters:

- *Relevancy* This parameter determines the level of relationship between a text and all texts, and causes the removal of those texts that are not sufficiently related to the texts. Equation 1 is used for calculating the relevancy of a text.

$$\text{Relevancy}T_i = \frac{n\left(\left(T_i\text{Concepts}\right) \cap (\text{ConceptsofallTexts})\right)}{n(T_i Concepts)} \tag{1}$$

- *Redundancy* This parameter determines the redundancy of a text in the produced summary and causes the removal of those texts that result in high redundancy. Indeed, the information of the texts with high redundancy there is in other texts and must be pruned. Equation 2 is used for calculating the redundancy of a text.

$$\text{Redundancy}T_i = \text{Max}\left(\frac{n\left(\left(T_i\text{Concepts}\right) \cap \left(T_j\text{Concepts}\right)\right)}{n\left(T_i\text{Concepts}\right)}\right) \tag{2}$$

$$\forall j = 1 \rightarrow n, j \neq i, n\left(T_j\text{Concepts}\right) \geq n(T_i\text{Concepts})$$

- *Coverage* This parameter determines the amount of information in other texts in a text and causes the removal of those texts that have low coverage on other texts. Equation 3 is used for calculating the coverage level of a text.

$$\text{Coverage}T_i = \frac{n\left(\left(T_i\text{Concepts}\right) \cap (\text{ConceptsofallTexts})\right)}{n(\text{ConceptsofallTexts})} \tag{3}$$

According to the pruning parameters, first, the relevancy parameter should be applied so that the texts with low relevancy are removed and the more relevant texts entered the process of other pruning parameters. In the following, according to the two parameters of redundancy and coverage, there are two possible arrangements (redundancy then coverage and coverage then redundancy). So, in the general case, the order of relevancy, coverage, and redundancy and the order of relevancy, redundancy, and coverage is applied.

The pseudocode for the multi-objective pruning of the texts in a cluster is shown in Algorithm 1. As can be seen in this algorithm, in order to prune the texts in the cluster, the text concepts in the texts are first extracted using the YAKE algorithm. Then, the texts are pruned iteratively. During pruning, the relevancy of the texts is first calculated. Afterward, a number of texts with low relevancy will be removed from the texts in the cluster. Thereafter, the redundancy of the texts will be calculated, removing those with high redundancy from the texts in the cluster. Finally, the coverage of the texts will be computed, where some texts with low coverage will be removed. This procedure continues until the number of remaining texts in the cluster reaches a predefined threshold. In this case, the texts remaining inside the cluster will be considered as the final summary of that cluster. Afterward, the summaries of the clusters are combined, the texts are again pruned based on the above procedure, and the final summary is obtained.

Algorithm 1: The pseudocode of the pruning approach based on concepts

```
Input: Opinion texts of a Cluster

Output: Summary of the cluster

Algorithm:

Extract concepts of the opinion texts by YAKE algorithm

Repeat

        Calculate the Relevancy of opinion texts

        Eliminate a number of opinion texts with minimum Relevancy

        Calculate the Coverage of opinion texts

        Eliminate a number of opinion texts with minimum Coverage

        Calculate the Redundancy of opinion texts

        Eliminate a number of opinion texts with maximum Redundancy

Until The number of the remained opinion texts is satisfied
```

Primary texts (Before Preprocessing):
- in the united states, #COVID cases are rising, but the death rate is relatively lo☺ .
- the virus deadlinessssss was overstated in the first place.
- It Is Just Bad Flu.
- We are getting close to herd immmmmmunity.
- physical distancing is making our immune systems weaker.
- #covid is caused or exacerbated BY 5g.
- masks do more harm than gooooood.
- doctors can already cure #COVID.
- big pharma is withholding the vacccccine.
- ANTIVIRALS and STEROIDS can cure COVID and CYTOKINE storms.

(a) Before Preprocessing

Primary texts (After Prepreocessing):
- in the united states, covid cases are rising, but the death rate is relatively low.
- the virus deadliness was overstated in the first place.
- it is just bad flu.
- we are getting close to herd immunity.
- physical distancing is making our immune systems weaker.
- covid is caused or exacerbated by 5g.
- masks do more harm than good.
- doctors can already cure covid.
- big pharma is withholding the vaccine.
- antivirals and steroids can cure covid and cytokine storms.

(b) After Preprocessing

Fig. 5 The initial opinion texts for summarization

Cluster: 0
- in the united states, covid cases are rising, but the death rate is relatively low.
- it is just bad flu.
- physical distancing is making our immune systems weaker.
- covid is caused or exacerbated by 5g.
- masks do more harm than good.

Cluster: 1
- the virus deadliness was overstated in the first place.
- we are getting close to herd immunity.
- doctors can already cure covid.
- big pharma is withholding the vaccine.
- antivirals and steroids can cure covid and cytokine storms.

Fig. 6 Opinion texts clustering

3.2 An Example of the proposed method

In order to show the details of the proposed method, 10 short opinion texts about COVID-19 were collected and stored in the form of a dataset. Then, the proposed method is applied to this dataset. As the first step, the dataset is read and the texts are preprocessed. Figure 5 indicates the 10 collected opinion texts after applying a number of preprocessing steps.

As explained regarding the steps of the proposed method, after performing preprocessing on the opinion texts, the dimensions of the texts must be reduced, and clustering must be performed. In this example, two clusters were considered. The results of the texts clustering are revealed in Fig. 6.

After clustering, the texts in each cluster are summarized using the multi-objective pruning approach based on relevancy, redundancy, and coverage parameters to obtain a summary for each cluster. The results of performing summarization on

Cluster: 0

~~- in the united states, covid cases are rising, but the death rate is relatively low.~~
~~- it is just bad flu.~~
- physical distancing is making our immune systems weaker.
~~- covid is caused or exacerbated by 5g.~~
- masks do more harm than good.

Cluster: 1

~~- the virus deadliness was overstated in the first place.~~
~~- we are getting close to herd immunity.~~
~~- doctors can already cure covid.~~
- big pharma is withholding the vaccine.
- antivirals and steroids can cure covid and cytokine storms.

Fig. 7 Summary of clusters

Final Summary:

~~- physical distancing is making our immune systems weaker.~~
- masks do more harm than good.
- big pharma is withholding the vaccine.
- antivirals and steroids can cure covid and cytokine storms.

Fig. 8 Final summary of 10 opinion texts

the clusters are shown in Fig. 7, where the pruned texts from each cluster are also specified.

Finally, once the summaries of the clusters are obtained, the multi-objective pruning approach is applied to the summaries of the two clusters to obtain the final summary. In this example, the length of the final summary is set to 3–4 texts. The final summary is shown in Fig. 8. Further, the texts pruned in the course of producing the final summary are specified in the figure.

4 Evaluating the efficiency of the proposed method

In order to show the efficiency of the proposed method, a number of simulations were performed. The simulations and evaluations are done in Python, only statistical t-Test is done in SPSS. In the following, the utilized datasets, the evaluation measures, and the results of the simulations are presented.

4.1 The used datasets

In order to measure the efficiency of the proposed method, standard and diverse datasets should be used to completely challenge the proposed approach. To do so, Sandy Hook Elementary School shooting in the USA (SH), Uttarakhand's floods (UK), Typhoon Hangupit in the Philippines (TH), and Bomb blasts in Hyderabad (HB) datasets were selected [33]. The SH, UK, TH, and HB datasets were collected from the tweets related to the four corresponding events and they included 2080,

Table 1 Details of the used datasets

Dataset name	Number of tweets	Number of tweets in gold summary
Sandy Hook elementary school shooting in USA (SH)	2080	37
Uttarakhand's floods (UK)	2069	34
Typhoon Hangupit in Philippines(TH)	1461	41
Bomb blasts in Hyderabad (HB)	1413	33

2069, 1461, and 1413 tweets, respectively. Three reference/gold summaries are presented along with each dataset to be used for evaluating the efficiency of the summarization methods. The number of tweets in the gold summaries for the SH, UK, TH, and HB datasets is 37, 34, 41, and 33 tweets, respectively. The details of the utilized datasets are reported in Table 1. These datasets can be downloaded for free from the link (http://crisisnlp.qcri.org/lrec2016/lrec2016.html).

4.2 The evaluation measures

One of the important methods for measuring the information level of the produced summary is the ROGUE[1] tool. The ROGUE tool includes a software package and a set of metrics that measure the extent of shared information between the summary created automatically and the summary created manually [44, 48, 52]. In order to determine the extent of shared information, the words present in the produced summary and those in the reference summary are compared in single-word, two-word, and multi-word manners, presented in the output as ROGUE-1, ROGUE-2, and ROGUE-L measures, respectively.

- ROUGE-1 It calculates the unigram overlap between the produced summary and the gold/reference summary.
- ROUGE-2 It calculates the bigram overlap between the produced summary and the gold/reference summary.
- ROUGE-L It calculates the longest common sequence overlap between the produced summary and the gold/reference summary.

The recall values are given in the results because, unlike precision, it measures the percentage of matching occurrences in both the produced summary and the gold/reference summary. The summary with the highest ROGUE value will be considered the summary closest to the real summary.

[1] Recall-Oriented Understudy for Gisting Evaluation.

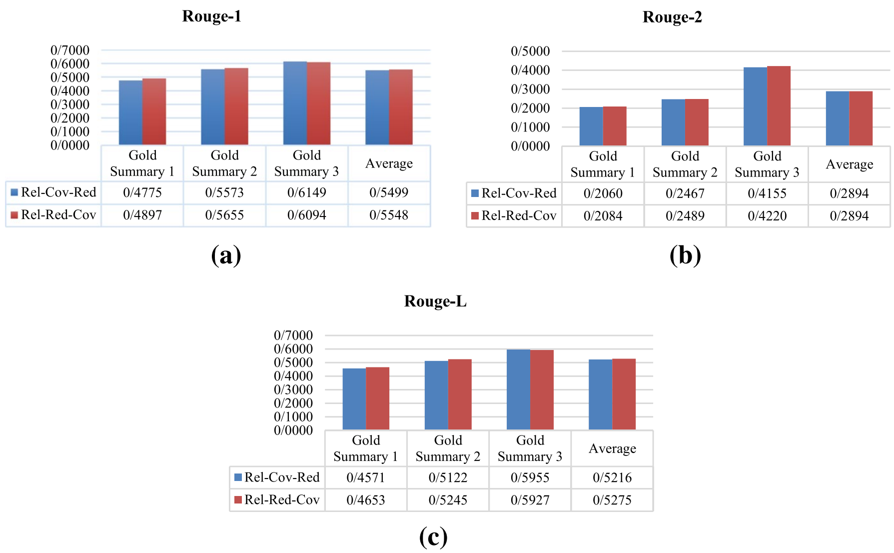

Fig. 9 Simulation results on HB dataset

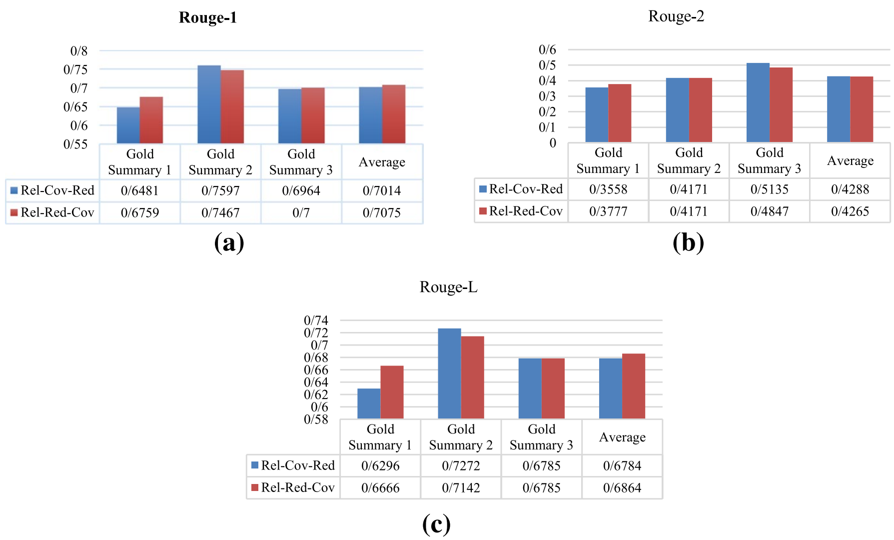

Fig. 10 Simulation results on SH dataset

4.3 Results

Figures 9, 10, 11 and 12 present the results of the simulations on the four selected datasets. The results are in the form of percentages and in the range of 0–1. In these simulations, the pruning parameters are once applied with the order of

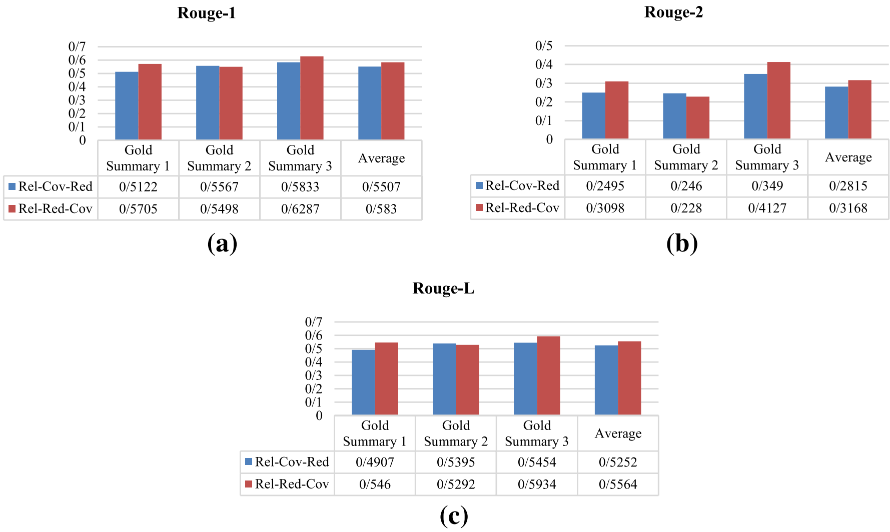

Fig. 11 Simulation results on TH dataset

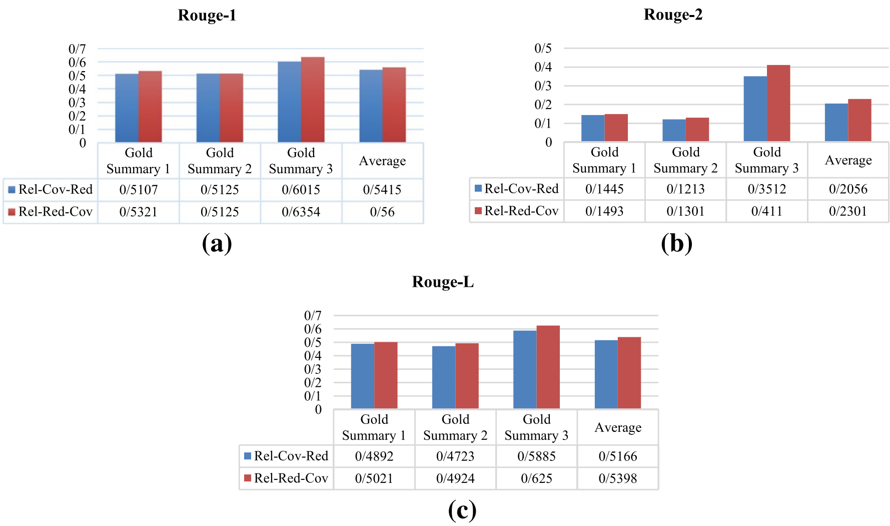

Fig. 12 Simulation results on UK dataset

Relevancy (Rel), Redundancy (Red), and Coverage (Cov), and once with the order of Relevancy (Rel), Coverage (Cov), and Redundancy (Red).

Figure 9 indicates the results of the simulations on the HB dataset. As can be seen in this figure, the results are shown in both scenarios, i.e., when pruning is performed with the order of Rel, Cov, and Red and when pruning is done with the order of Rel, Red, and Cov. The results related to ROGUE-1, ROGUE-2, and

Table 2 Simulation results on 4 datasets

Dataset	Proposed method (Rel-Cov-Red)			Proposed method (Rel-Red-Cov)			MOOTweetSumm		
	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
HB	0.5499	0.2894	0.5216	0.5548	0.2931	0.5275	0.5371	0.3914	0.5371
SH	0.7014	0.4288	0.6784	0.7075	0.4265	0.6864	0.5842	0.3721	0.5842
TH	0.5507	0.2815	0.5252	0.5830	0.3168	0.5564	0.3845	0.2184	0.3782
UK	0.5415	0.2056	0.5166	0.5600	0.2301	0.5398	0.4541	0.2822	0.4447

ROGUE-L measures are separately shown based on the gold summaries. Furthermore, the average results related to the gold summaries are also presented. As can be observed, when pruning is performed with the order of Rel, Red, and Cov, better results have been obtained.

Figure 10 reveals the simulation results on the SH dataset. The results for ROGUE-1, ROGUE-2, and ROGUE-L measures are displayed based on both directions of pruning. For this dataset, ROGUE-1 and ROGUE-L measures had a better performance when the pruning was performed with the order of Rel-Red-Cov; however, the ROGUE-2 measure performed better with the pruning order of Rel-Cov-Red.

Figure 11 illustrates the simulation results for the TH dataset. As can be seen, for this dataset, all the measures had a better performance in summarizing opinion texts when the pruning was performed with the order of Rel-Red-Cov.

Figure 12 reveals the simulation results for the UK dataset based on ROGUE-1, ROGUE-2, and ROGUE-L parameters. In this dataset, both pruning directions of Rel-Red-Cov and Rel-Cov-Red were applied with the results for all three measures presented. Similar to HB and TH datasets, for this dataset, the pruning direction of Rel-Red-Cov resulted in a better summarization in terms of all three measures.

Table 2 presents the results obtained from the proposed method compared to the MOOTweetSumm method [33] on the four datasets, where better results are shown in bold. Since the MOOTweetSumm method performs best among the state-of-the-art methods, the proposed method was compared to this method. Furthermore, the MOOTweetSumm method has different modes and this table presents the best value obtained from these modes.

In order to better compare the proposed method to the MOOTweetSumm method, Fig. 13 depicts the average simulation results on the four datasets using the proposed method in both Rel-Red-Cov and Rel-Cov-Red directions compared to all modes of the MOOTweetSumm method. As can be seen from the figure, in terms of the average results on the four datasets, the proposed method has provided the best summarization performance in the Rel-Red-Cov direction.

Table 3 reports the average simulation results for the four datasets compared to the state-of-the-art methods in terms of ROGUE-2 and ROGUE-L criteria. As can be seen, the proposed method has outperformed the state-of-the-art methods.

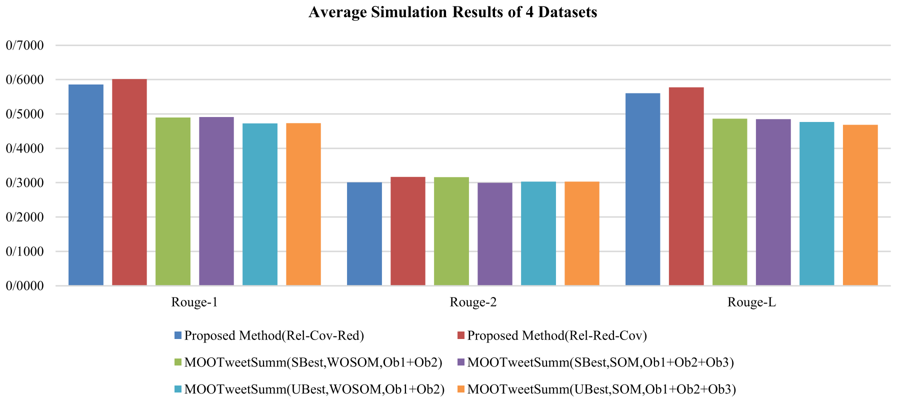

Fig. 13 Simulation results compared with MOOTweetSumm method

Table 3 Simulation results compared with state-of-the-art methods

Method	Rouge-2	Rouge-L
Proposed method (Rel-Red-Cov)	0.3166	0.5775
Proposed method (Rel-Cov-Red)	0.3013	0.5604
MOOTweetSumm (SBest, WOSOM, Ob1 + Ob2) [33]	0.3160	0.4860
MOOTweetSumm (SBest, SOM, Ob1 + Ob2 + Ob3) [33]	0.2999	0.4850
MOOTweetSumm (UBest, WOSOM, Ob1 + Ob2) [33]	0.3033	0.4769
MOOTweetSumm (UBest, SOM, Ob1 + Ob2 + Ob3)[33]	0.3033	0.4681
VecSim-ConComp-MaxDeg [22]	0.1919	0.4457
VecSim-ConComp-MaxLen [22]	0.1940	0.4506
VecSim-ConComp-maxSumTFIDF [22]	0.1886	0.4600
VecSim-Community-maxSumTFIDF [22]	0.1898	0.4591
ClusterRank (CR) [23]	0.0859	0.2684
COWTS (CW) [21]	0.1790	0.4454
Lex-Rank (LR) [24]	0.0489	0.1525
LSA (LS) [25]	0.1599	0.4234
LUHN (LH) [26]	0.1650	0.4015
Mead (MD) [27]	0.1172	0.3709
SumBasic (SB) [28]	0.1012	0.3289
SumDSDR (SM) [29]	0.0985	0.2602

The reason behind reporting ROGUE-2 and ROGUE-L in this table is that the papers related to previous methods have reported the results for these measures.

4.4 Statistical t-test

To validate the results obtained by the proposed method, a statistical significance t-Test is done in SPSS software. It is carried out to check whether the average

ROUGE scores (Table 3) obtained by the proposed approach are statistically significant or occurred randomly. The output of the t-Test is p-value and the smaller p-value confirms that our results are significant. The p-values obtained using Table 3 are 0.00003 using the ROUGE-2 score and 0.015 using the ROUGE-L score. Test results confirm that obtained results by the proposed method didn't occur randomly and improvements are statistically significant.

4.5 Discussion and evaluation of the results

As can be seen from Figs. 9, 10, 11 and 12, when pruning is performed with the order of Rel, Red, and Cov, better results are obtained. The reason behind the better performance with this direction of pruning is that texts with lower relevancy are first pruned and more relevant texts go to the next stage. Then, among the remaining texts, those resulting in higher redundancy are pruned, while the texts resulting in lower redundancy in the texts present in the summarization are transferred to the next step. Finally, when more relevant texts with lower redundancy are transferred to the next step, the texts resulting in lower coverage are pruned, thus enhancing the coverage of the summarized texts. In addition, the results in Figs. 9, 10, 11 and 12 indicate that the ROGUE-2 measure has a lower value compared to the ROGUE-1 and ROGUE-L measures. This discrepancy is caused by the fact that in the ROGUE-1 measure, the comparison between the produced summary and the gold summary is made word-by-word, while the comparison for the ROGUE-L measure is done in a multi-word format, resulting in higher values for this measure. However, the comparison for the ROGUE-2 measure is made in a two-word manner, and since two-word phrases present in the produced summary can be highly different from those in the gold summary, this measure will have a lower value.

Based on the results presented in Tables 2 and 3 along with Fig. 13, the proposed method has had a better performance in opinion text summarization. Based on the obtained results, the proposed method has provided an 11% improvement in terms of the ROGUE-1 measure compared to previous methods. The improvement obtained in terms of the ROGUE-2 measure is trivial. Further, an improvement of about 9% has been obtained in terms of the ROGUE-L measure, indicating the acceptable performance of the proposed approach. The reason behind the insignificant improvement in the ROGUE-2 measure is that in the proposed method, the concepts are considered individually and generally in the pruning parameters; however, there is no pairwise evaluation of the concepts. As a result, there is only a slight improvement in terms of the ROGUE-2 measure.

As noted earlier, the main contribution of the proposed method is to reduce the dimension of opinion texts before clustering and produce summaries based on the concepts present in the original texts using the multi-objective pruning approach based on Relevancy, Redundancy, and Coverage Parameters. The reason behind the better performance of the proposed approach is that before clustering, the dimension of the opinion texts is reduced. Further, after identifying the intrinsic dimensions of the texts, the clustering process is performed, improving the clustering performance, as confirmed in [53] while previous works use base clustering algorithms. Producing accurate clusters has a direct impact on the result of the summarization. Secondly, when producing the summary, the concepts in the original texts are identified and the summarization is performed based on these identified concepts, which also has a direct influence on the summarization result, while in previous works, summarization based on concepts isn't done. In addition, when producing the summary, instead of a selection mechanism, a pruning mechanism based on coverage, redundancy, and relevancy parameters is utilized. The pruning mechanism gradually removes texts with lower importance, while texts of higher importance are added to the final summary. In addition, the parameters used in the pruning process are the parameters that directly improve the quality of the final summary. In fact, instead of single-step selection, less important texts are gradually removed from the set of texts, if such an approach wasn't used in previous works.

5 Conclusions

The use of various social networks has significantly increased in recent years where people with different social backgrounds express their sentiments and opinions regarding various issues in the form of short texts. These sentiments and opinions are a great decision-making source for other individuals as they are used in different areas. Considering the very large volume of these texts, it is not easy for everyone to analyze and make use of these texts. As a result, automatic analysis of these opinion texts, especially the summarization of opinion texts, can be a great help in different areas. When summarizing opinion texts, the goal is to receive a set of short texts as well as produce a comprehensive and useful summary of the opinions and useful information of these texts. In this paper, a new approach is presented for summarizing opinion texts. The proposed approach performs the summarization of opinion texts using manifold learning and the concepts in these texts. Accordingly, in order to overcome the challenge of the complexity of opinion texts, the dimensions of these texts are first reduced using manifold learning. Then, the concepts present in the texts are used as a basis to select important texts to be added to the produced summary using the multi-objective pruning method. In order to generate summaries with high quality, pruning was performed based on relevancy, redundancy, and coverage parameter. Two different pruning directions based on Rel-Red-Cov and Rel-Cov-Red were applied to the proposed method. The simulation results revealed that the Rel-Red-Cov pruning direction could provide better performance for the summarization of opinion texts. Further, the proposed method outperformed state-of-the-art methods in terms of ROGUE-1, ROGUE-2, and ROGUE-L measures.

The limitations of this work are performing pruning as static and improvement in ROUGE-2 measure is trivial. Future works in this area can follow two different approaches. In the first approach, the pruned texts can be evaluated and if the pruned text can be added to the final summary, it can again be added to the set of selected texts, also we can do pruning as dynamic. In addition, since there was no significant improvement in terms of the ROGUE-2 measure, this issue can be considered when formulating the pruning parameters in an attempt to reach a significant improvement in terms of this parameter as well.

Acknowledgements Not applicable.

Author's contribution All authors contributed to the ideation, theoretical development, and design of the research framework. SJ carried out the experiment under the supervision of AMEM and FM. All authors discussed the results and contributed to the final manuscript. Also, all authors confirm the results and the Manuscript content.

Funding Not applicable.

Data availability The data used to support the findings of this study are available from the corresponding author upon request.

Declarations

Conflict of interest The authors declare that there are no competing interests regarding the publication of this paper.

Consent for publication We consent to the publication of our paper in the Journal of Supercomputing.

References

1. Pabbi K, Sindhu C (2021) Opinion summarisation using bi-directional long-short term memory. Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET).
2. Murray G, Hoque E, Carenini G (2017) Sentiment analysis in social networks. Morgan Kaufmann
3. Aamir M, Ullah Jan A, Mukhtar N, Abid Khan M, Ali Z, Ahmed Abro W, Yurong G (2021) An unsupervised graph-based hybrid approach for opinion summarization. 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP).
4. Varma V, Kurisinkel LJ, Radhakrishnan P (2017) A practical guide to sentiment analysis. Springer
5. Saeed RM, Rady S, Gharib TF (2021) Optimizing sentiment classification for Arabic opinion texts. Cogn Comput 13:164–178
6. Saeed NMK, Helal NA, Badr NL, Gharib TF (2019) An enhanced feature-based sentiment analysis approach. Wiley Interdiscipl Rev: Data Min Knowl Discov 10:1–20
7. Golande A, Kamble R, Waghere S (2016) An overview of feature based opinion mining. Intell Syst Technol Appl 530:633–645
8. Moussa ME, Mohamed EH, Haggag MH (2018) A survey on opinion summarization techniques for social media. Fut Comput Inf J 3:82–109
9. Nathania G, Siautama R, Claire A, Suhartono D (2021) Extractive hotel review summarization based on TF/IDF and adjective-noun pairing by considering annual sentiment trends. Procedia Computer Sci 179:558–565
10. Moratanch N, Chitrakala S (2017) A survey on extractive text summarization. IEEE International Conference on Computer, Communication, and Signal Processing (ICCCSP-2017).
11. Moratanch N, Chitrakala S (2016) A survey on abstractive text summarization. International Conference on Circuit, Power and Computing Technologies [ICCPCT].
12. Almatrafi Q, Johri A (2022) Improving MOOCs using information fromdiscussion forums: an opinion summarization and suggestion mining approach. IEEE Access 10:15565–15573
13. Huang Y, Yu Zh, Xiang Y, Yu Zh, Guo J (2022) Exploiting comments information to improve legal public opinion news abstractive summarization. Front Comp Sci 16:1–10
14. Rani R, Lobiyal DK (2022) Document vector embedding based extractive text summarization system for Hindi and English text. Appl Intell 25:840
15. Modi Sh, Oza R (2018) Review on abstractive text summarization techniques (ATST) for single and multi documents. International Conference on Computing, Power and Communication Technologies (GUCON): 1173–1176.
16. Naqeeb Khan S, Mohd Nawi N, Imrona M, Shahzad A, Ullah A, Rahman A (2018) Opinion mining summarization and automation process: a survey. Int J Adv Sci Eng Inf Technol 8:1836–1844

17. Gupta S, Gupta SK (2019) Abstractive summarization: an overview of the state of the art. Expert Syst Appl 121:49–65
18. Ouyang L, Li W, LiLu SQ (2018) Applying regression models to query-focused multi-document summarization. Inf Process Manag 47:227–237
19. Mahak G, Vishal G (2017) Recent automatic text summarization techniques: a survey. Artif Intell Rev 48:1–66
20. Khan A, Salim N, Kuma YJ (2015) A framework for multi-document abstractive summarization based on semantic role labeling. Appl Soft Comput 30:737–747
21. Rudra K, Ghosh S, Ganguly N, Goyal P, Ghosh S (2015) Extracting situational information from microblogs during disaster events: A classification-summarization approach. Proceedings of the 24th ACM International on Conference on Information and Knowledge Management: 583–592.
22. Dutta S, Chandra V, Mehra K, Das AK, Chakraborty T, Ghosh S (2018) Ensemble algorithms for microblog summarization. IEEE Intell Syst 3:4–14
23. Garg N, Favre B, Reidhammer K, Hakkani-Tur D (2009) ClusterRank: A graph based method for meeting summarization. Proceedings of the 10th International Conference of the International Speech Communication Association (Interspeech 2009): 1499–1502.
24. Erkan G, Radev DR (2004) LexRank: graph-based lexical centrality as salience in text summarization. J Artif Intell 22:457–479
25. Gong Y, Liu X (2001) Generic text summarization using relevance measure and latent semantic analysis. Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval: 19–25.
26. Luhn HP (1958) The automatic creation of literature abstracts. IBM J Res Dev 2:159–165
27. Radev DR, Hovy E, McKeown K (2002) Introduction to the special issue on summarization. Comput Linguist 28:399–440
28. Nenkova A, Vanderwende L (2005) The impact of frequency on summarization. Microsoft Res. 101:87
29. He Z et al. (2012) Document summarization based on data reconstruction. Twenty-Sixth AAAI Conference on Artificial Intelligence: 620–626.
30. Dutta S, Chandra V, Mehra K, Ghatak S, Das AK, Ghosh S (2019) Summarizing microblogs during emergency events: a comparison of extractive summarization algorithms. Emerg Technol Data Min Inf Secur 45:859–872
31. Alguliyev RM, Aluliyev RM, Isazadeh NR, Abdi A, Idris N (2016) COSUM: text summarization based on clustering and optimization. Expert Syst 6:1–17
32. Chakraborty R, Bhavsar M, Kumar Dandapat S, Chandra J (2019) Tweet summarization of news articles: an objective ordering-based perspective. IEEE Trans Comput Soc Syst 6:761–777
33. Saini N, Saha S, Bhattacharyya P (2019) Multiobjective-based approach for microblog summarization. IEEE Trans Comput Soc Syst 6:1219–1231
34. Dutta S, Das AK, Bhattacharya A, Dutta G, Parikh KK, Das A, Ganguly D (2019) Community detection based tweet summarization. emerging technologies in data mining and information security. Adv Intell Syst Comput 813:797–808
35. Zhou X, Wan X, Xiao J (2016) CMiner: opinion extraction and summarization for chinese microblogs. IEEE Trans Knowl Data Eng 28:1650–1663
36. Jebrakumar R, Shrivastava S, Ramchandani H (2018) Multi post summarization using clustering technique. Proceedings of 4th International Conference on Energy Efficient Technologies for Sustainability.
37. Niu J, Zhao Q, Wang L, Chen H, Zheng Sh (2016) Opinion summarization for short texts based on BM25 and syntactic parsing. 14th International Conference on Industrial Informatics (INDIN): 1177–1180.
38. Waheeb SA, Khan NA, Chen B, Shang X (2020) Multidocument arabic text summarization based on clustering and word2vec to reduce redundancy. Information 11:1–13
39. Yu C, Shu L (2019) The Summarization of commodity short comments based on topic clustering. Journal of Physics: Conference Series 1302.
40. Loret E, Boldrini E, Vodolazova T, Martínez-Barco P, Muñoz R, Palomar M (2015) A novel concept-level approach for ultra-concise opinion summarization. Expert Syst Appl 42:7148–7156
41. Amplayo RK, Song M (2017) An adaptable fine-grained sentiment analysis for summarization of multiple short online reviews. Data Knowl Eng 110:54–67

42. Bhatia S, Madan R, Yadav S L, Bhatia K (2019) An algoritmic approach based on principal component analysis for aspect-based opinion summarization. 6th International Conference on Computmohding for Sustainable Global Development (INDIACom): 874–879.
43. Roul RK, Sahoo JK (2020) Sentiment analysis and extractive summarization based recommendation system. Comput Intell Data Min 990:473–487
44. Abdi C, Shasuddin SM, Hasan S (2018) Machine learning-based multi-documents sentiment-oriented summarization using linguistic treatment. Expert Syst Appl 109:66–85
45. Ma Y, Li Q (2019) A weakly-supervised extractive framework for sentiment-preserving document summarization. World Wide Web 22:1401–1425
46. Zhen Y, Fei Y, Kefeng F, Jian H (2017) Text dimensionality reduction with mutual information preserving mapping. Chin J Electron 26:919–925
47. Pozzi FA, Fersini E, Messina E, Liu B (2017) Sentiment analysis in social networks. Morgan Kaufmann
48. Lopez Gondori RE, Salaguerio Pardo TR (2017) Opinion summarization methods: comparing and extending extractive and abstractive approaches. Expert Syst Appl 7:124–134
49. Rehioui H, Idrissi A (2020) New clustering algorithms for twitter sentiment analysis. IEEE Syst J 14:530–537
50. Shih-Hung L, Kuan-Yu Ch, Berlin Ch, Hsin-Min W, Wen-Lian H (2017) Leveraging manifold learning for extractive broadcast news summarization. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): 5805–5809.
51. Campos R, Mangaravite V, Pasquali A, Jorge A, Nunes C, Jatowt A (2020) YAKE! Keyword extraction from single documents using multiple local features. Inf Sci 509:257–289
52. Wang R, Luo S, Pan L, Wu Zh, Yuan Y, Chen Q (2019) Microblog summarization using paragraph vector and semantic structure. Comput Speech Lang 57:1–19
53. Jahanbakhsh Gudakahriz S, Eftekhari Moghadam AM, Mahmoudi F (2021) Opinion texts clustering using manifold learning based on sentiment and semantics analysis. Scientif Program 2021:5

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.