A Hybrid Ensemble Word Embedding based Classification Model for Multi-document Summarization Process on Large Multi-domain Document Sets

S Anjali Devi, S Sivakumar
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Guntur, Andhra Pradesh, India -522502

***Abstract*—Contextual text feature extraction and classification play a vital role in the multi-document summarization process. Natural language processing (NLP) is one of the essential text mining tools which is used to preprocess and analyze the large document sets. Most of the conventional single document feature extraction measures are independent of contextual relationships among the different contextual feature sets for the document categorization process. Also, these conventional word embedding models such as TF-ID, ITF-ID and Glove are difficult to integrate into the multi-domain feature extraction and classification process due to a high misclassification rate and large candidate sets. To address these concerns, an advanced multi-document summarization framework was developed and tested on number of large training datasets. In this work, a hybrid multi-domain glove word embedding model, multi-document clustering and classification model were implemented to improve the multi-document summarization process for multi-domain document sets. Experimental results prove that the proposed multi-document summarization approach has improved efficiency in terms of accuracy, precision, recall, F-score and run time (ms) than the existing models.**

Keywords*—Word embedding models; text classification; multi-document summarization; contextual feature similarity; natural language processing

I. Introduction

Machine learning (ML) has become a key approach to problem solving and data predictions. Machine learning allows a classifier to learn a set of rules, or the criterion of decision, from a set of labelled data that an expert has annotated. This approach enables better scaling and reduced time when classifying topic domain data as compared to a system that relies only on manual input. Most of the research was done on binary classifiers in the field of machine learning based on the classification of multi-domain document data. In many fields, the purpose of using machine learning for pattern mining has an important role in decision-making systems. A set of input documents is split into two or more classes in the text classification (TC) process [1], with each document belonging to one or more classes depending on its contents. Document clustering [2] is the method of categorizing text documents into a hierarchical cluster or category, so that the documents are identical in the same cluster, whereas the documents in the other clusters are different. It is one of the vitals of text mining processes. In particular, text mining has gained significant significance and involves various tasks, such as the development of granular taxonomies, document summarization, etc., to produce knowledge of higher quality from text. The supervised strategy is utilized to solve the problem if we have a predetermined class or classes. A prediction-based model is a decision tree. It is distinguished by a tree-like system of rules and is mostly used to solve classification problems. The decision tree is built using data from training. With this strategy, a tree is built to represent the categorization problem. The majority of previous works [3] used single-document summarization. Approaches based on sentence extraction from documents are used in single-document summarization. Most single-document summarization systems employ a simple method for summary generation, which consists of extracting the first sentence from each paragraph and placing them in the same order as they were written. Later on, the presence of multiple sources delivering the same information causes problems for news providers' end users, who must read the same material repeatedly. As a result, recent work [4] has centred on multi-document summarization. To combine information held in distinct documents for multi-document summarization, valuable procedures are necessary. This usually means that some operations, such as key matches, matching terms, sentence position, and sentence length, must be performed below the sentence level. As a summary [5], multi-document summarization may successfully handle the concerns by generating shorter summaries including the important points of the original documents using criteria for decreasing redundancy and maximising variety in the selected articles. Before reordering the phrases into the document's original sequence, most extractive summary optimisation algorithms score them based on their value. Without access to the real summary analysis mechanism, it is not always possible to build partial rank lists of sentences using only the original document and the summary. The two major types of text summarization are abstraction and extraction [6]. The sentence with the highest score among the other sentences is chosen during the document extraction process. Whereas, abstraction entails employing linguistic techniques to create something new, which may or may not be present in the source, and substituting it for the summary without altering its original meaning. The entire collection is searched for important objects in the extractive summarization task, with no changes to the objects themselves. Conciseness, accuracy, and objectivity are three qualities of a good summarizer. The goal of this paper's proposed methodology [7] is to create an extractive text summarizer that can generate variable-length summaries. According to [8], the summary frequently includes sentences that are not closely related to one another. This can be handled by generating the sentence set with a sufficient threshold. As a result, one of our issues is deciding on a sufficient threshold. The order of the sentences in the summary is the next problem. Another challenge with news summarization systems is how to handle huge feature sets, as the complexity of weight adjustment increases exponentially as the number of features increases. As a result, higher-performance systems with more useful features are required. Among the three types of summarization systems, extractive summarization is perhaps the most investigated. Although the phrase is most commonly used to refer to sentence extraction and reordering, numerous extractive approaches also focus on sub-sentence extraction. An extractive system can be topic-based, centrality-based, or a combination of the two. The relevance of particular words or phrases is prioritized in topic-based systems. Although specialized machine learning techniques such as neural networks (NN) and support vector machines (SVM) are used in many fields to classify data into one or more classes, traditional models must be improved on large datasets with high dimensionality. Some demonstrations of supervised learning include Linear Regression, Logistic Regressions, Decision Trees, and SVM. These are some demonstrations of supervised learning. Classification [9] can be defined as the procedure of classifying objects of interest into different previously defined categories or classes.

Recently, extractive single document summaries have been generated using machine learning methods. Nave-bayes, Hidden Markov Model (HMM), and log-linear models are some of the methodologies that fall within machine learning approaches. Automatic text summarization using artificial intelligence and neural networks has been the subject of a few studies. Given a set of features, the Hidden Markov model [HMM] estimates [10] the posterior probability that each phrase is a summary sentence or not. This model has fewer assumptions of independence than the naive Bayesian approach. The number of terms in the sentence, as well as the likelihood of the terms given the baseline of terms (Baseline term probability) [11] and the document terms (Document term probability).

Wrapper techniques use a black box for a single learner to evaluate the function subsets on the basis of their predictive effectiveness. The embedded techniques select the features in the integrated phase and are generally particular to one individual instance. PSO and neural action provide a possible optimization solution [12]. Each particle accelerates during each iteration towards the best global location discovered by the representative points. Scalability is inefficient at identifying the globally optimal solution. Dynamic goals and connectivity are taken as tasks rather than restrictions. The Multi-Objective Data Relations (MODP) approach is used to resolve all existing problems in order to improve anomaly relations. Further work can be undertaken in the future in order to significantly reduce the normalized root-mean-square error. Recently, ensemble learning models have become popular and widely accepted for high-dimensional and imbalanced datasets. Most of the traditional ensemble classification models are processed with limited feature space and small data size. As the size of the feature space increases, traditional ensemble classifiers select a predefined number of features for classification. The main objective of the ensemble learning models based on feature selection is to classify high-dimensional features on high-dimensional datasets with high computational efficiency and a high true positive rate [13]. Severe problems such as performance and scalability may result from learning classification models with all their high-dimensional features. Many textual content classifiers [14] have been proposed in the literature, including those that use machine learning techniques, probabilistic models, and so on. Decision trees, nave-bayes, rule induction, neural networks, nearest [15-17], and, most recently, guide vector machines are some of the techniques used.

The main contributions in this paper are:

1) Proposed a hybrid multi-domain glove optimization model on the large document sets.

2) Proposed a multi-document clustering method for the document summarization process.

3) Implemented a hybrid multi-document Bayesian approach based document summarization process on large document sets.

The main sections presented in this paper are:

Section II describes the overall literature work of the word embedding models and multi-document summarization. In the section III, a hybrid word embedding measures are proposed in order to classify the multi-domain features for the multi-document summarization process. Also, a hybrid multi-document cluster based classification model is proposed in the section 3. In the section IV, experimental results and its discussion are discussed. Finally, in the section V, conclusion of the work is presented.

II. Related Work

Wu et.al, proposed key extraction by combining multidimensional information, and they named their proposed system MIKE. They used two datasets from the ACM world wide web to form the ACM knowledge discovery and data mining. They compared their results to the TF-IDF and TextRank algorithms to assess their performance [18]. LAKE is a key phrase based summarizer system that extracts relevant key phrases from documents using statistical analysis. In terms of text summarization methods, neural networks outperform other traditional methods in terms of extractive methods for handling semantics and redundancy, but fall short in terms of coherence when compared to abstractive methods. There are various approaches to abstractive summarization, including linguistic-based approaches, semantic graph-based approaches, and hybrid extractive/abstractive approaches [19]. Syntactic representations and tree structures are used in linguistic-based approaches, but semantic meanings are not abstracted. As discussed in a previous study, semantic graph-based approaches focus on semantic role labelling to determine the abstraction of input to core meaning to form graphs to filter out redundancy, followed by a text generator to build summaries. Extractive methods are used in hybrid approaches to obtain an output summary that is fed into a text generator to build non-key words and phrases to improve sentence coherence and readability.

SUMMARIST [20] is a key phrase summarizer used to find the boundaries of extraction using a rank-based abstraction approach. The FEMsum summarizer is used to create summaries using a graph-representation and to identify the relationships between the candidate sentences, as well as a syntactic and semantic representation of the phrases. The data structure required for recognizingtopics in document sets and creating various forms of summaries is built by using a fuzzy co-reference cluster graph technique [21]. The intra- and inter-document co-reference chain families generated by a co-reference method under various (fuzzy) clustering criteria are given as input to this algorithm. In other words, each cluster assigns a topic to each document: some themes appear in all documents (common topic), while others appear in only a subset or a single document (contrastive/distinctive topic). In [22], a set of distance functions for assessing structural similarity between online documents is analyzed. They analysed different Tag Frequency Distribution Analysis (TFDA), parametric functions, and edit distance between documents as three distinct ways of defining similarity. [23] proposed a label discovery technique that uses a hierarchical structure to express the relationship between text data in online documents collected from the web. Their programme correctly classifies web pages by discovering similar labels that describe the same type of content. [24] utilised a model that combined documents from various taxonomies. For the classification challenge, their model used the Nave Bayes algorithm. Content-based classifiers are clearly used by some research tools, such as NewsDude, to select valuable articles and to remove articles that appear to be excessively repetitious of previously read articles. [25] proposed employing a support vector machine (SVM) classifier to identify web pages based on both text and context features. They tested their online classification methods using the WebKB dataset, and the results demonstrate that using context features, particularly hyperlinks, can greatly enhance classification performance.

Conventional statistical methods have been included in many models. The main drawback of conventional statistical methods is the rigidity of dynamic situations and therefore the difficulty of optimal modelling. Most of the traditional ensemble classification models are processed with limited feature space and small data size. As the size of the feature space increases, traditional ensemble classifiers select a predefined number of features for classification. [26] proposed a novel discretization approach to continuous attributes for decision tree learning. The main issue with traditional decision tree models is that each attribute is assumed to be either nominal or categorical. To overcome this issue, a dynamic discretization model on the continuous label is applied to each attribute during the tree construction process. Traditional decision tree models such as CART and C4.5 use discretization methods in the preprocessing phase along with noise removal methods. But, the main limitation of this model is that the data should be of a continuous type and it doesn't support mixed types.

Feature selection is a process that selects an optimal feature sub-set based on a particular requirement. The measuring feature subsets are specified in the criterion. The criterion will be selected according to the purposes for which the feature is selected. For example, an optimal subset can be a minimum subset. It can provide the best predictive accuracy estimate in a sub-set. In some circumstances [27], a subset with the specified number that meets the criterion can be found in view of the number of features. Rough Sets Attribute Reduction (RSAR) is a filter-based tool for feature reduction used to extract data and maintain information while reducing the amount of knowledge involved. Analysis of Rough Sets is performed on the basis only of the data provided, and no external parameters are required to operate [28]. This makes use of the data granularity structure. It does, however, continue to assume the model that there is some information available with every item in the discourse universe that truly and accurately reflects the real world. The ideal criterion for the selection of Rough Sets is to find the shortest or minimum reductions while obtaining high grades for the selected features. The redundancy of a feature or feature subset is determined. A feature is declared relevant if the decision feature is predictive, otherwise it is irrelevant. A Principal Component Analysis (PCA) approach to a reduction in dimension is achieved by building main components that are linear combinations of the original predictor or the explaining variables. The PCA approach is based on the supposition that large variance in characteristics provides useful information, and, in contrast, small variance is considered less useful. Ortholy-linear combinations have been designed to maximize features in the linear combination of explicative variables. There are two basic stages of Fuzzy ELM (F-ELM), [29] called preparation and prediction. P. Verma and H. Om [30] proposed the Correlation-based Feature Selection (CFS) method. The correlation between the attribute and the class is calculated by this approach, with the hypothesis that an optimal collection of features should be strongly correlated with the class but not correlated with other features. This is to ensure that redundancies and feature numbers [31-34] (explaining the pattern with as few features as possible but still maintaining high performance) are reduced. Artificial intelligence is a notion that today has a lot of excitement around it. They trained the decision tree using a rotated feature space. Hence, they proposed the rotation forest algorithm. In this method,[35-37]samples from the main datasets are obtained. These samples form a new subset which is fed into a new feature space.

III. Proposed Model

Initially, document sets are taken as input for text pre-processing. In the preprocessing phase, each document is preprocessing using the Stanford NLP library. This library is used to perform various operations such as document tokenization, stemming and stop word removal on different domain fields. After performing the data pre-processing operations on the large document sets, word embedding model is used to optimize the document to word vectors. In this work, a hybrid multi-domain word embedding model is proposed in order to optimize the word embedding key words on large document-sets.

Proposed multi-domain glove optimization model is designed to find the main and its contextual key features on large document sets. Multi-document contextual features are extracted using the main words of the glove model. A boosting contextual similarity is computed based of the main words, contextual words, string hash similarity and multi-document contextual similarity features to filter essential top k voted features in the document sets. In the next step, a multi-document clustering approach is developed on the filtered top k-contextual voted features for the multi-document summarization process. In the multi-document clustering process, an efficient KNN distance measure is used to compute the nearest clusters by using the structural similarity between the main and contextual scores. Each document and its key features are labelled with the cluster class for the multi-document summarization process. In the proposed multi-document summarization, a hybrid Bayesian probability based classification approach is developed to find the multi-document summarization process as shown in Fig.1.

A. *Multi-Document Glove Optimization Word Embedding Model*

In the multi-document glove optimization model, each pre-processed document is given as input to compute word co-occurrence matric. Let X_{ij} represents the word occurrence matrix in order to compute main word and contextual word on large document set. W_i and W_j represent main and contextual word vectors of $W_{c.}$

$b_i, b_j : main\ \ word\ bias\ vector\ \ and\ context\ word\ bias\ vector$
$\theta = Min\{b_i, b_j\}. D(\,(w_i . b_i), (w_j\,, b_j\,))$

1) Defining multi-document summarization cost function and its constraints using the word, main vectors and its biases:

$$C = CostFunction = tan^{-1}(\eta) * b_i w_i^T w_j + b_j w_i^T w_j + exp(\eta) * max\{\,b_m, b_c\} - \frac{cos(\,X_{ij}}{\sqrt[3]{||w_i|| * \left||w_j|\right|}}$$

$multi - document$ weights are defined as

$$\eta = multi_{weight} = f\left(X_{ij}\right) = \begin{cases} (\frac{\sqrt{X_{ij}}}{X_{max}})^{\alpha} & \text{if } X_{ij} < X_{MAX} \\ \sqrt{X_{ij}}^{\alpha} & \text{otherwise} \end{cases}$$

$where\ \alpha$ is scaling factor.

2) Define a multi-document cost function.

$$\text{J}=\sum_{\text{i=1}}^{V}\sum_{\text{j=1}}^{V} \eta.\,(tan^{-1}(\,\theta) * b_i w_i^T w_j + b_j w_i^T w_j + exp(\,\theta) * max\{\,b_m, b_c\} - (cos(\,X_{ij})/\sqrt[3]{||w_i|| * ||w_j||})^2$$

3) The Proposed Multi-document word embedding model is optimized by using the partial derivative w.r.t main words and contextual words as shown below.

$$\frac{\partial \text{J}}{\partial \text{w}_\text{i}} = \text{b}_\text{i}\text{w}_\text{j}\,\text{C} = \text{b}_\text{i}\text{w}_\text{j}(\tan^{-1}(\theta) * b_i w_i^T w_j + b_j w_i^T w_j + \exp(\theta) * \max(b_m, b_c) - (\cos X_{ij})/\sqrt[3]{\left||w_i|\right| * \left||w_j|\right|}$$

$$\frac{\partial \text{J}}{\partial \text{w}_\text{j}} = \text{b}_\text{i}\text{w}_\text{i}\,\text{C} = \text{b}_\text{i}\text{w}_\text{i}(\tan^{-1}(\theta) * b_i w_i^T w_j + b_j w_i^T w_j + \exp(\theta) * \max(b_m, b_c) - (\cos X_{ij})/\sqrt[3]{\left||w_i|\right| * \left||w_j|\right|}$$

update w_i and w_j using learning theta.

In the above optimized multi-domain glove optimization model, the cost function and its constraints are improved in order to find the essential key contextual features among the multiple domain document sets. Here, the multiweight factor is used to find the weighted document features among the main and contextual feature vectors. Finally, the multi-document cost function is based on multi-weights, main and contextual feature vectors on large contextual co-occurrence matrix.

B. *Boosting Voting based Word Embedding Contextual Similarity*

In this phase, a voted boosting method is used to compute the best similarity measure based on the multi-document glove main and contextual key vectors. In this phase, hash based similarity, string similarity and proposed multi-document main and contextual similarity measure are used to choose the majority voted similarity on the glove main and contextual feature vectors. The proposed main and contextual similarity measure is computed by using the following formula.

$let\ \omega(i)$ be the multi-document main word vector features ,

$\omega(j) represents$ the multi-document contextual
word vector features.

$$\chi = \frac{cos(\,\omega(i), \omega(j)) \times \sum_{i=1}^{k}\left(\omega(i) - \overline{\omega(i)}\right)}{\sqrt{\sum_{i=1}^{k}\left|\omega(i) - \overline{\omega(i)}\right| \times \sum_{j=1}^{k}\left|\omega(j) - \overline{\omega(j)}\right|}}$$

$$Multi - document \text{ main word similarity} = tf(i) * log(\,\chi/\,max\{\,|\omega(i)|, |\omega(j)|)$$

$$Multi - document \text{ contextual word similarity} = l\,og(\,1 + \chi/\,min\{\,|\omega(i)|, |\omega(j)|)$$

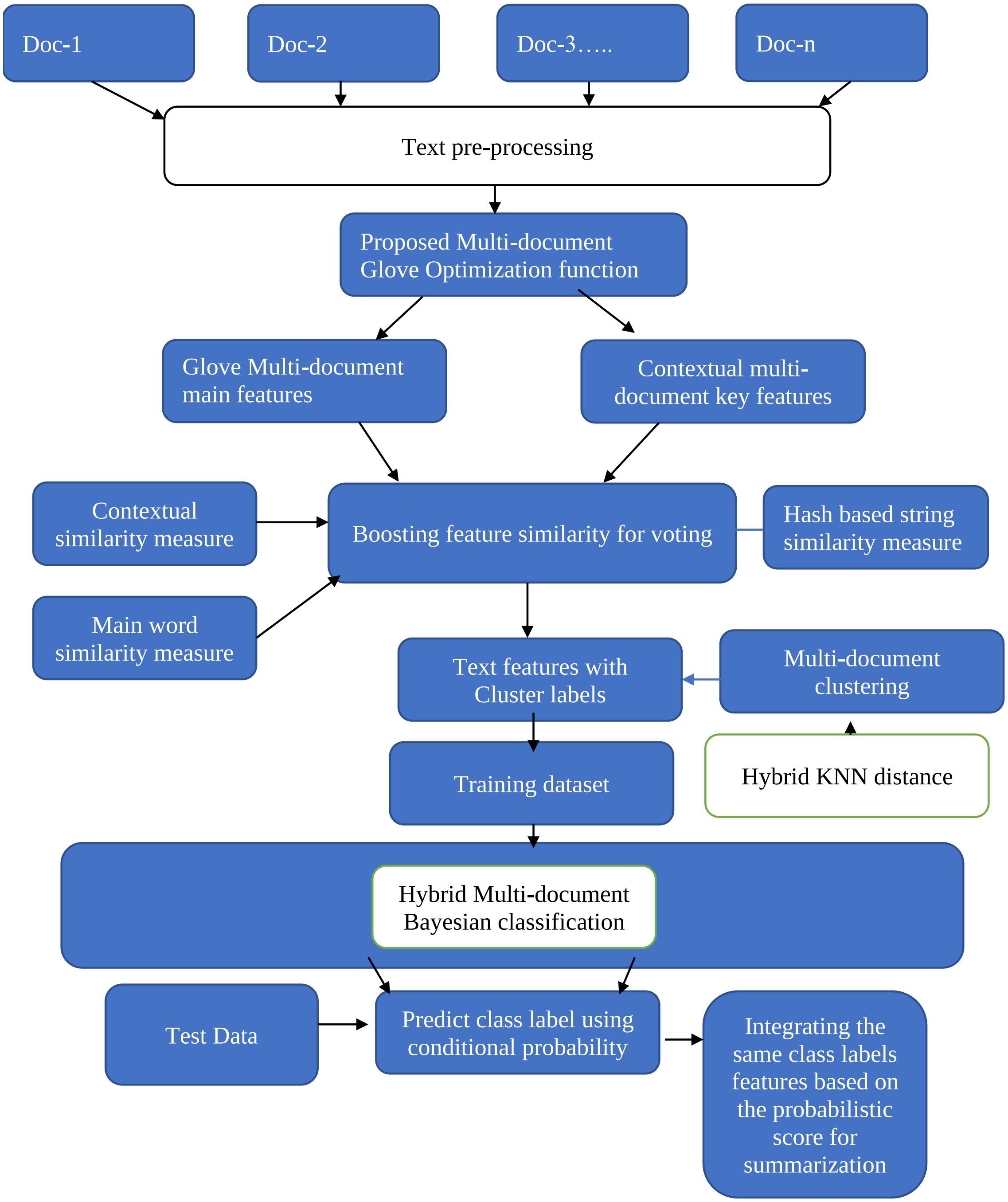
Doc-1
Doc-2
Doc-3…..
Doc-n
Text pre-processing
Proposed Multi-document Glove Optimization function
Glove Multi-document main features
Contextual multi-document key features
Contextual similarity measure
Boosting feature similarity for voting
Hash based string similarity measure
Main word similarity measure
Text features with Cluster labels
Multi-document clustering
Training dataset
Hybrid KNN distance
Hybrid Multi-document Bayesian classification
Test Data
Predict class label using conditional probability
Integrating the same class labels features based on the probabilistic score for summarization

Fig. 1. Proposed Multi-Domain Multi-Document Summarization Framework.

C. Multi-Document Clustering based on KNN Approach

In this phase, a hybrid multi-document clustering based KNN approach is developed on the main and contextual key similarity features. This approach is used to group the multi-documents based on the domain main and contextual similarity vectors. Let k defines the user defined number of k-nearest objects for grouping.

let MD_t represents the multi-document sets.

Output: Clustered k documents with cluster class label

Procedure:

1) Read input data MD_t

2) Initialize k clusters for KNN and perform traditional k-means document clustering algorithm.

3) In the proposed document clustering approach, instead of using the conventional distance measures, a hybrid weighted distance measure is proposed between the main and contextual word vectors.

4) The weighted multi-document pair distance between the main and contextual word vectors is given as

$$\psi_M(TF_{t,d}) = P(w_m / D_i).\frac{tf_{t,d} \times \log \frac{\sqrt{\chi}}{|w_m|}}{[(\sum_{t=1}^{n}(1+\log(\frac{|D|}{\sqrt{T_m}})]}$$

$$\psi_C(TF_{t,d}) = P(w_c / D_i).\frac{tf_{t,d} \times \log \frac{\sqrt{\chi}}{|w_c|}}{[(\sum_{t=1}^{n}(1+\log(\frac{|D|}{\sqrt{T_c}})]}$$

Where

$$\text{where } tf_{t,d}, \eta > 0$$

and

$$T_c = \sum_{d=1}^{D}\sum_{t} tf_{t,d}, T_m = \sum_{d=1}^{D}\sum_{t} tf_{t,d}$$

$$\text{Weightedpairscore(WPS)} = \sum |\psi_M(TF_{t,d}) - \psi_C(TF_{t,d})|$$

Finally, the contextual similarity between the main word vectors and contextual word vectors are clustered using the following similarity score measure as

$$S(MW(d_i), CW(d_j)_j) = \psi_M(TF_{t,d}).\frac{\sum_{k=1}^{n} t_{ik} \times t_{jk}}{\sqrt{\sum_{k=1}^{n} t_{ik}^2}\sqrt{\sum_{k=1}^{n} t_{jk}^2}} + \psi_C(TF_{t,d}).\frac{\sum_{k=1}^{n} t_{ik} \times t_{jk}}{\sqrt{\sum_{k=1}^{n} t_{ik}^2}\sqrt{\sum_{k=1}^{n} t_{jk}^2}}$$

5) The kscore is used to find the document classification score in each domain filed for the class label prediction on the new test data. The kscore measure is computed using the following formula.

$$\text{KScore } (D_t, C_k) = \sum_{d \in DK} S(MW(d_i), CW(d_j)_j) \times P(D_i, C_j)$$

$$P(D_i, C_j) = \begin{cases} 1 & D_i \in Cj \\ 0 & D_i \notin C_j \end{cases}$$

D. Multi-Document Conditional Bayesian Estimation based Classification

In the multi-document summarization phase, the clustered training data which is generated in the previous section are taken as input to the multi-document base multi-domain classification process. Proposed Bayesian classification model is used to classify the key phrases for the multi-document summarization process. In this phase, two optimizations are performed on the traditional Bayesian text classification model. In the first optimization, a hybrid prior multi-document probability is developed to predict the multi-domain phase on the large textual document sets. In the second optimization, a hybrid posterior probability is proposed on the main and contextual word vectors in each class category. The main steps used in the proposed multi-document summarization are

1) Read contextual and main words clustered labelled document sets as input.

2) Compute prior multi-document classification probability as:

$$\begin{aligned} Pr(MW(d_i), CW(d_j)) &= Multi - Doc((MW(d_i), CW(d_j), C(k)) \\ &= P(MW(d_i)/ C(k)) \\ &* max\{P(CW \cap MW) / C(k)\}/|MW(d_i) \\ &+ CW(d_j)| \end{aligned}$$

3) Predict the posterior multi-document estimation using the maximization of the class labels as:

$$\begin{aligned} ClassPredict(MW(d_i), CW(d_j)) &= argmax\{P(C(k)) \\ &* \{\prod P(CW \cup MW) / C(k)\}/ ||D| \\ &- (MW(d_i) + CW(d_j))| \end{aligned}$$

4) To each document in the training documents sets,Merge the phrase with high posterior probability and contextual-main word similarity scores for summarization process.

IV. Experimental Results

The performance has been evaluated using the multi-document summarization datasets provided by Document Understanding Conferences (DUC) 2002, Document Understanding Conferences (DUC) 2004[38], multi-news [39], multi-biomedical datasets [40]. It is an open benchmark from the National Institute of Standards and Technology (NIST) for the evaluation of generic automatic summarization. The experiments have been carried out in amazon AWS server with 96 GB RAM.

In the experimental study, word embedding features, classification metrics and multi-document summarization rouge metrics are used to evaluate the performance of the proposed model to the conventional models.

Table 1, illustrates the performance evaluation of the proposed multi-domain glove optimization model to the conventional approaches for contextual similarity computation on DUC 2002 dataset. As represented in the table, the proposed approach has improved evaluation than the previous models in terms of contextual similarity between the main and contextual word vectors.

Figure 2, illustrates the performance evaluation of the proposed multi-domain glove optimization model to the conventional approaches for contextual similarity computation on DUC 2004 dataset. As represented in the table, the proposed approach has improved evaluation than the previous models in terms of contextual similarity between the main and contextual word vectors.

TABLE I. COMPARATIVE ANALYSIS OF PROPOSED MODEL TO CONVENTIONAL MODELS FOR OVERALL CONTEXTUAL SIMILARITY MEASURE (DUC:2002)

TestDoc	MI	Entropy	TextRank	Glove	Proposed Glove
TestCat1-1	0.86	0.89	0.92	0.93	0.96
TestCat1-2	0.88	0.9	0.91	0.93	0.95
TestCat1-3	0.87	0.89	0.92	0.94	0.95
TestCat1-4	0.86	0.89	0.9	0.93	0.96
TestCat1-5	0.87	0.89	0.92	0.93	0.95
TestCat1-6	0.87	0.89	0.92	0.94	0.95
TestCat1-7	0.88	0.9	0.91	0.92	0.95
TestCat1-8	0.86	0.91	0.91	0.93	0.95
TestCat1-9	0.89	0.89	0.92	0.93	0.95
TestCat1-10	0.85	0.88	0.92	0.93	0.95
TestCat1-11	0.86	0.88	0.91	0.92	0.95
TestCat1-12	0.86	0.9	0.91	0.93	0.95
TestCat1-13	0.85	0.88	0.91	0.92	0.97
TestCat1-14	0.86	0.91	0.92	0.93	0.97
TestCat1-15	0.85	0.9	0.92	0.93	0.96
TestCat1-16	0.88	0.89	0.91	0.94	0.95
TestCat1-17	0.86	0.88	0.92	0.94	0.96
TestCat1-18	0.85	0.88	0.91	0.94	0.96
TestCat1-19	0.88	0.89	0.9	0.94	0.96
TestCat1-20	0.87	0.89	0.91	0.93	0.95

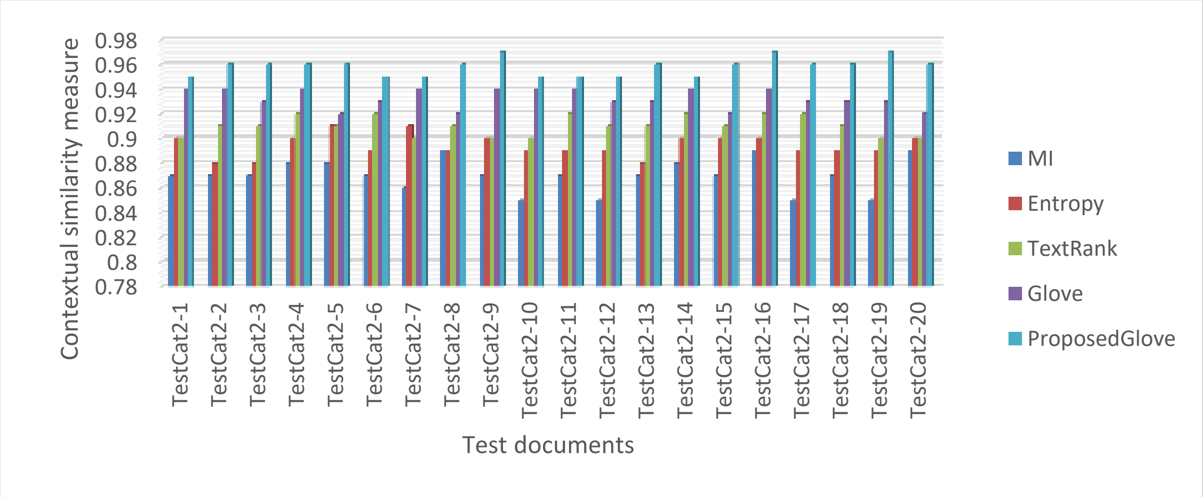

Fig. 2. Comparative Analysis of Proposed Model to Conventional Models for Overall Contextual Similarity Measure (DUC:2004).

Table 2, illustrates the performance evaluation of the proposed multi-domain glove optimization model to the conventional approaches for contextual similarity computation on multi-news dataset. As represented in the table, the proposed approach has improved evaluation than the previous models in terms of contextual similarity between the main and contextual word vectors.

Table 3, illustrates the performance evaluation of the proposed multi-domain glove optimization model to the conventional approaches for contextual similarity computation on multi-biomedical dataset. As represented in the table, the proposed approach has improved evaluation than the previous models in terms of contextual similarity between the main and contextual word vectors.

Figure 3, illustrates the performance evaluation of the proposed multi-domain glove optimization model to the conventional approaches for filtering optimal key features count for document clustering process on DUC 2002 dataset. As represented in the table, the proposed approach has improved evaluation than the previous models in terms of contextual keywords filtering among the large number of candidate key features space.

TABLE II. COMPARATIVE ANALYSIS OF PROPOSED MODEL TO CONVENTIONAL MODELS FOR OVERALL CONTEXTUAL SIMILARITY MEASURE (MULTI-NEWS)

TestDoc	MI	Entropy	TextRank	Glove	ProposedGlove
TestCat3-1	0.86	0.88	0.91	0.92	0.95
TestCat3-2	0.88	0.91	0.92	0.94	0.95
TestCat3-3	0.87	0.89	0.9	0.94	0.95
TestCat3-4	0.86	0.88	0.91	0.93	0.95
TestCat3-5	0.85	0.89	0.9	0.93	0.96
TestCat3-6	0.88	0.88	0.92	0.94	0.95
TestCat3-7	0.89	0.9	0.92	0.92	0.95
TestCat3-8	0.88	0.88	0.92	0.94	0.96
TestCat3-9	0.89	0.9	0.92	0.93	0.95
TestCat3-10	0.87	0.9	0.9	0.93	0.95
TestCat3-11	0.88	0.89	0.91	0.94	0.96
TestCat3-12	0.87	0.9	0.92	0.93	0.95
TestCat3-13	0.87	0.9	0.9	0.94	0.95
TestCat3-14	0.88	0.91	0.91	0.92	0.96
TestCat3-15	0.85	0.89	0.9	0.94	0.97
TestCat3-16	0.87	0.9	0.9	0.94	0.95
TestCat3-17	0.89	0.89	0.9	0.94	0.96
TestCat3-18	0.88	0.89	0.91	0.94	0.95
TestCat3-19	0.87	0.9	0.91	0.93	0.97
TestCat3-20	0.88	0.88	0.92	0.94	0.95

TABLE III. COMPARATIVE ANALYSIS OF PROPOSED MODEL TO CONVENTIONAL MODELS FOR OVERALL CONTEXTUAL SIMILARITY MEASURE (BIOMEDICAL DOCS)

TestDoc	MI	Entropy	TextRank	Glove	ProposedGlove
TestCat4-1	0.87	0.89	0.91	0.92	0.95
TestCat4-2	0.86	0.88	0.91	0.94	0.96
TestCat4-3	0.86	0.89	0.92	0.92	0.95
TestCat4-4	0.85	0.89	0.92	0.93	0.95
TestCat4-5	0.85	0.9	0.92	0.93	0.95
TestCat4-6	0.89	0.88	0.92	0.92	0.95
TestCat4-7	0.88	0.9	0.92	0.93	0.96
TestCat4-8	0.85	0.9	0.91	0.94	0.95
TestCat4-9	0.89	0.89	0.91	0.92	0.96
TestCat4-10	0.85	0.9	0.92	0.94	0.95
TestCat4-11	0.88	0.91	0.92	0.94	0.96
TestCat4-12	0.88	0.88	0.92	0.93	0.95
TestCat4-13	0.85	0.89	0.91	0.93	0.97
TestCat4-14	0.89	0.89	0.92	0.92	0.96
TestCat4-15	0.86	0.9	0.9	0.93	0.95
TestCat4-16	0.89	0.89	0.91	0.92	0.95
TestCat4-17	0.89	0.89	0.92	0.94	0.95
TestCat4-18	0.86	0.89	0.9	0.93	0.97
TestCat4-19	0.85	0.88	0.91	0.94	0.96
TestCat4-20	0.88	0.89	0.92	0.93	0.95

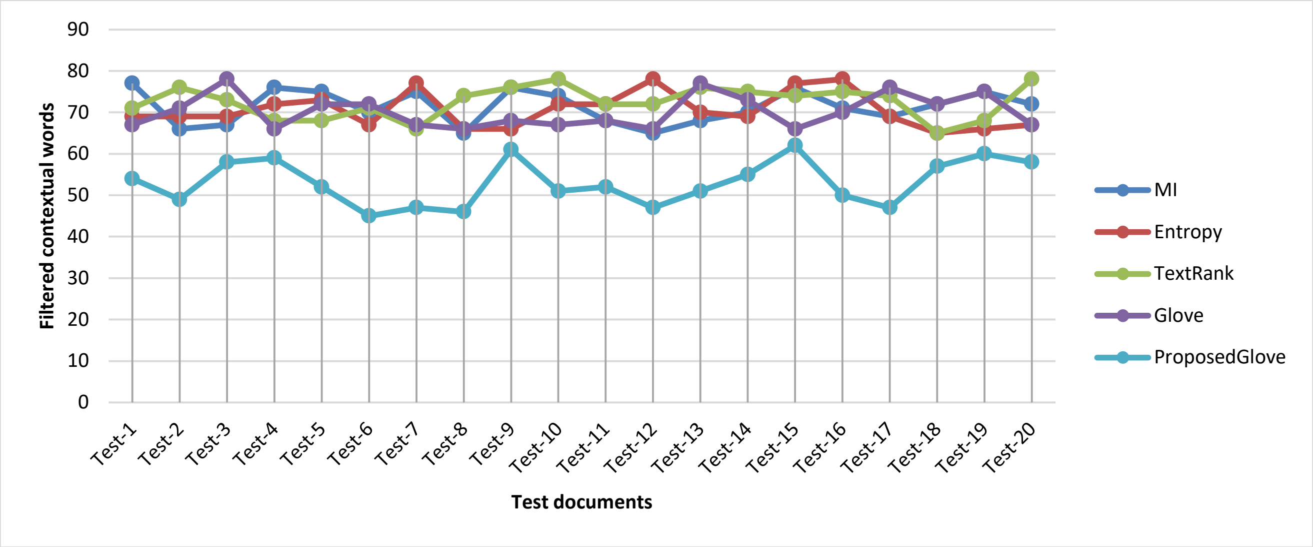

Fig. 3. Comparative Analysis of Proposed Model to Conventional Models for Overall Contextual Keywords Filtering (DUC 2002).

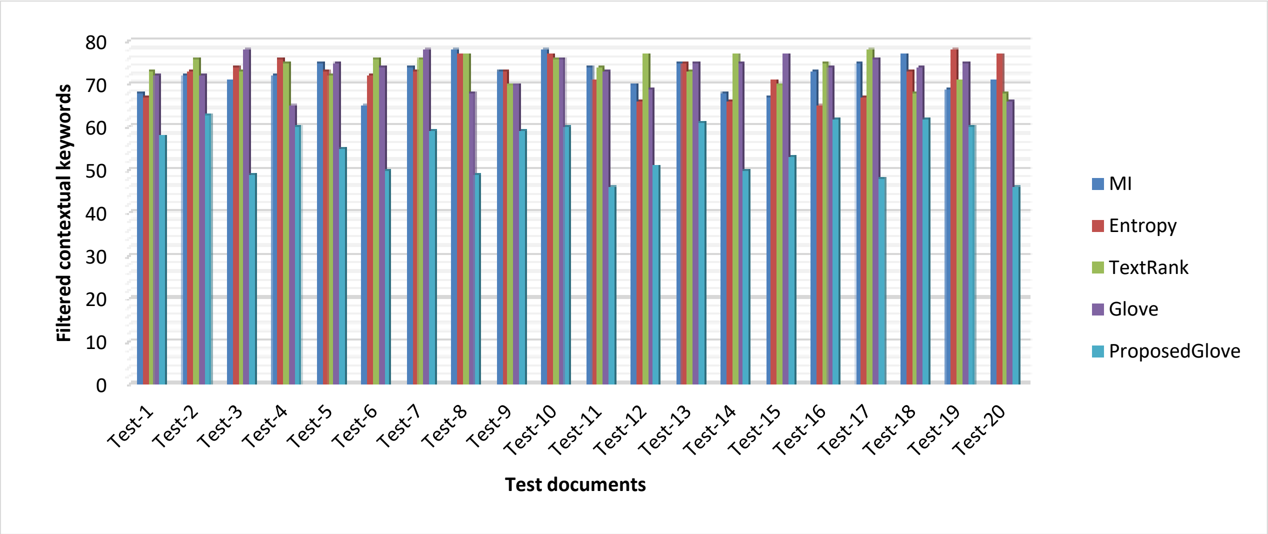

Fig. 4. Comparative Analysis of Proposed Model to Conventional Models for Overall Contextual Keywords Filtering (DUC 2004).

Figure 4, illustrates the performance evaluation of the proposed multi-domain glove optimization model to the conventional approaches for filtering optimal key features count for document clustering process on DUC 2004 dataset. From the figure, it is observed that the proposed approach has improved evaluation than the previous models in terms of contextual keywords filtering among the large number of candidate key features space.

Table 4, illustrates the performance evaluation of the proposed multi-domain glove optimization model to the conventional approaches for filtering optimal key features count for document clustering process on multi-news dataset. As represented in the table, the proposed approach has improved evaluation than the previous models in terms of contextual keywords filtering among the large number of candidate key features space.

Table 5, illustrates the performance evaluation of the proposed multi-domain glove optimization model to the conventional approaches for filtering optimal key features count for document clustering process on biomedical document sets. As represented in the table, the proposed approach has improved evaluation than the previous models in terms of contextual keywords filtering among the large number of candidate key features space.

Table 6, represents the performance evaluation of the proposed multi-document based Bayesian summarization model to the conventional models for classification precision on various domain databases. From the table, it is noted that the proposed multi-document based Bayesian summarization approach has improved precision than the previous approaches on different domain document sets.

Figure 5, represents the performance evaluation of the proposed multi-document based Bayesian summarization model to the conventional models for classification accuracy on various domain databases. From the figure, it is noted that the proposed multi-document based Bayesian summarization approach has improved accuracy than the previous approaches on different domain document sets.

Figure 6, represents the performance evaluation of the proposed multi-document based Bayesian summarization model to the conventional models for classification recall on various domain databases. From the figure, it is noted that the proposed multi-document based Bayesian summarization approach has improved recall than the previous approaches on different domain document sets.

TABLE IV. COMPARATIVE ANALYSIS OF PROPOSED MODEL TO CONVENTIONAL MODELS FOR OVERALL CONTEXTUAL KEYWORDS FILTERING (MULTI-NEWS)

TestDoc	MI	Entropy	TextRank	Glove	ProposedGlove
Test-1	73	73	70	77	61
Test-2	67	70	68	69	53
Test-3	65	66	76	68	57
Test-4	77	76	71	68	45
Test-5	74	67	70	77	54
Test-6	78	69	73	76	59
Test-7	73	73	68	71	57
Test-8	74	73	68	73	50
Test-9	67	71	68	76	62
Test-10	67	65	74	71	53
Test-11	72	71	74	70	51
Test-12	77	77	73	70	49
Test-13	68	65	66	65	54
Test-14	69	71	76	67	53
Test-15	70	72	71	76	57
Test-16	66	77	72	74	57
Test-17	71	75	71	77	57
Test-18	74	69	70	69	50
Test-19	67	75	77	70	62
Test-20	77	65	74	65	55

TABLE V. COMPARATIVE ANALYSIS OF PROPOSED MODEL TO CONVENTIONAL MODELS FOR OVERALL CONTEXTUAL KEYWORDS FILTERING (BIOMEDICAL DOCS)

TestDoc	MI	Entropy	TextRank	Glove	ProposedGlove
Test-1	72	75	66	76	60
Test-2	69	78	71	77	57
Test-3	72	69	68	69	55
Test-4	77	67	76	69	57
Test-5	76	77	66	72	55
Test-6	73	73	72	75	52
Test-7	67	66	72	73	56
Test-8	70	73	67	71	54
Test-9	68	74	71	70	61
Test-10	72	65	77	73	56
Test-11	69	69	69	65	48
Test-12	72	67	75	66	62
Test-13	71	75	72	73	52
Test-14	66	73	65	66	54
Test-15	75	68	73	69	52
Test-16	68	73	72	75	46
Test-17	71	71	67	77	48
Test-18	66	75	68	76	50
Test-19	69	67	72	70	52
Test-20	72	73	67	72	53

TABLE VI. COMPARATIVE EVALUATION OF PROPOSED MULTI-DOCUMENT BASED BAYESIAN SUMMARIZATION MODEL TO THE CONVENTIONAL MODELS FOR CLASSIFICATION PRECISION ON VARIOUS DOMAIN DATABASES

MultiDoc Test	CSTSumm	GistSumm	MultiLayer	ProposedMDBayesian Summ
DUC2002	0.86	0.89	0.9	0.97
DUC2004	0.87	0.89	0.91	0.96
Multi-News	0.86	0.9	0.9	0.95
Multi-Biomedical	0.85	0.89	0.9	0.95

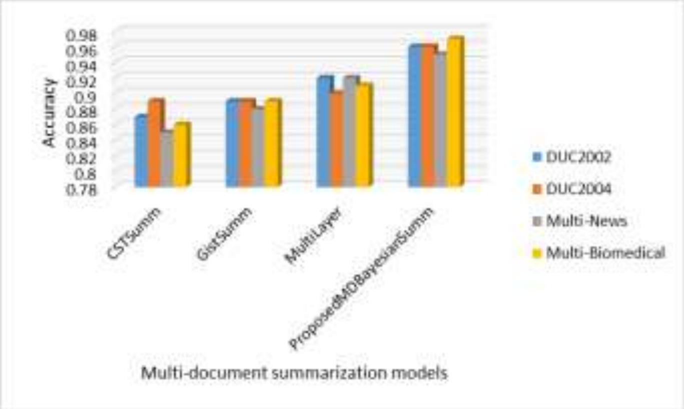

Fig. 5. Comparative Evaluation of Proposed Multi-Document based Bayesian Summarization Model to the Conventional Models for Classification Accuracy on Various Domain Databases.

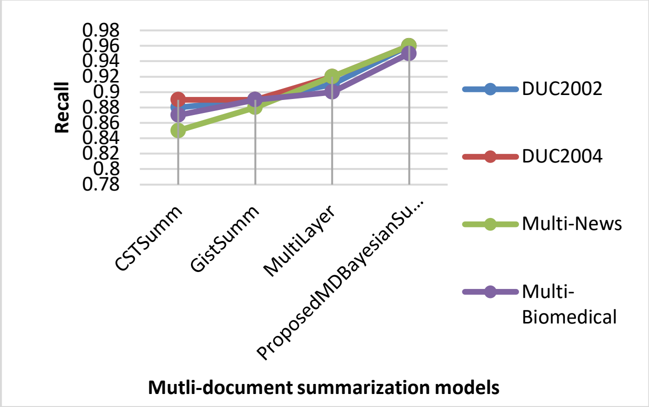

Fig. 6. Comparative Evaluation of Proposed Multi-Document based Bayesian Summarization Model to the Conventional Models for Classification Recall on Various Domain Databases.

For experimental evaluation, we use ROUGE (Recall-Oriented Understudy for Gisting Evaluation) in order to find the performance of the proposed multi-doc summarization process on various traditional models.

Table 7, represents the performance evaluation of the proposed multi-document based Bayesian summarization model to the conventional models for average rouge metrics on DUC 2002 domain database. From the table, it is noted that the proposed multi-document based Bayesian summarization approach has improved average rouge metrics than the previous approaches on different DUC 2002 document sets.Table 8, represents the performance evaluation of theproposed multi-document based Bayesian summarization model to the conventional models for average rouge metrics on multi-news domain database. From the table, it is noted that the proposed multi-document based Bayesian summarization approach has improved average rouge metrics than the previous approaches on different domain multi-news data.

TABLE VII. COMPARATIVE EVALUATION OF PROPOSED MULTI-DOCUMENT BASED BAYESIAN SUMMARIZATION MODEL TO THE CONVENTIONAL MODELS FOR AVERAGE ROUGE METRICS ON VARIOUS DUC 2002 DATABASE

Avg Rouge	Gensim	OPINOSIS	PyTextRank	Proposed BayesianSumm
Recall	0.05	0.065	0.087	0.17
Precision	0.04	0.075	0.075	0.14
F-measure	0.054	0.065	0.0734	0.12

TABLE VIII. COMPARATIVE EVALUATION OF PROPOSED MULTI-DOCUMENT BASED BAYESIAN SUMMARIZATION MODEL TO THE CONVENTIONAL MODELS FOR AVERAGE ROUGE METRICS ON MULTI-NEWS DOMAIN DATABASE

Avg Rouge	Gensim	OPINOSIS	PyTextRank	ProposedBayesianSumm
Recall	0.034	0.046	0.085	0.14
Precision	0.023	0.048	0.078	0.094
F-measure	0.036	0.05	0.09	0.12

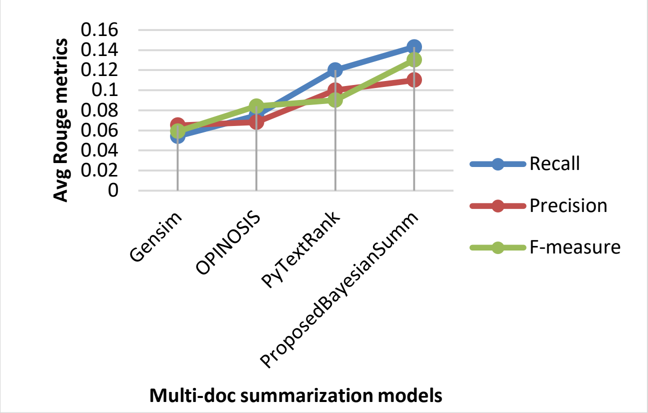

Fig. 7. Comparative Evaluation of Proposed Multi-Document based Bayesian Summarization Model to the Conventional Models for Average Rouge Metrics on Multi-Biomedical Domain Database.

Figure 7, represents the performance evaluation of the proposed multi-document based Bayesian summarization model to the conventional models for average rouge metrics on multi-biomedical domain database. From the table, it is noted that the proposed multi-document based Bayesian summarization approach has improved average rouge metrics than the previous approaches on biomedical document sets.

TABLE IX. COMPARATIVE EVALUATION OF PROPOSED MULTI-DOCUMENT BASED BAYESIAN SUMMARIZATION MODEL TO THE CONVENTIONAL MODELS FOR AVERAGE ROUGE METRICS ON VARIOUS DUC 2004 DATABASE

Avg Rouge	**Gensim**	**OPINOSIS**	**PyTextRank**	**ProposedBayesianSumm**
Recall	0.035	0.084	0.095	0.16
Precision	0.043	0.078	0.11	0.154
F-measure	0.049	0.069	0.12	0.158

Table 9, represents the performance evaluation of the proposed multi-document based Bayesian summarization model to the conventional models for average rouge metrics on DUC 2004 domain database. From the table, it is noted that the proposed multi-document based Bayesian summarization approach has improved average rouge metrics than the previous approaches on different DUC 2004 document sets.

A. Results Interpretation

In this work, different multi-document features and its correlated main and contextual words are used to analyze the multiple documents for summarization. From the experimental results it is noted that the average accuracy, recall and precision of the proposed multi-document summarization is better than the conventional models with nearly 1% improvement. Also, the contextual features of the proposed glove model has better optimization for the word to vector generation process.

V. CONCLUSION

Multi-document summarization plays a vital role in the multi-domain document sets due to variation in the feature space and inter and intra document cluster variations. Since, most of the conventional multi-document summarization models have large number of candidate feature sets for document clustering and classification process. In this work, a hybrid multi-document based glove optimization model is proposed in order to filter the key features on multi-domain document sets. Also, a hybrid document clustering and multi-document Bayesian classification model for multi-domain document summarization process is proposed on large document sets. Experimental evaluation represent the performance of the proposed Bayesian multi-document summarization approach has improved rouge evaluation metrics than the previous models with nearly 2-3% improvement on large multi-domain document sets. In the future scope, this work can be extended to improve the multi-level based dynamic multi-domain feature extraction and summarization process using the parallel processing framework.

REFERENCES

[1] J. M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Pérez, “A decomposition-based multi-objective optimization approach for extractive multi-document text summarization,” Applied Soft Computing, vol. 91, p. 106231, Jun. 2020, doi: 10.1016/j.asoc.2020.106231.

[2] A. Abdi, S. Hasan, S. M. Shamsuddin, N. Idris, and J. Piran, “A hybrid deep learning architecture for opinion-oriented multi-document summarization based on multi-feature fusion,” Knowledge-Based Systems, vol. 213, p. 106658, Feb. 2021, doi: 10.1016/j.knosys.2020.106658.

[3] R. Ferreira et al., “A multi-document summarization system based on statistics and linguistic treatment,” Expert Systems with Applications, vol. 41, no. 13, pp. 5780–5787, Oct. 2014, doi: 10.1016/j.eswa.2014.03.023.

[4] G. Yang, D. Wen, Kinshuk, N.-S. Chen, and E. Sutinen, “A novel contextual topic model for multi-document summarization,” Expert Systems with Applications, vol. 42, no. 3, pp. 1340–1352, Feb. 2015, doi: 10.1016/j.eswa.2014.09.015.

[5] M. Mojrian and S. A. Mirroshandel, “A novel extractive multi-document text summarization system using quantum-inspired genetic algorithm: MTSQIGA,” Expert Systems with Applications, vol. 171, p. 114555, Jun. 2021, doi: 10.1016/j.eswa.2020.114555.

[6] D. Bollegala, N. Okazaki, and M. Ishizuka, “A preference learning approach to sentence ordering for multi-document summarization,” Information Sciences, vol. 217, pp. 78–95, Dec. 2012, doi: 10.1016/j.ins.2012.06.015.

[7] M. Bidoki, M. R. Moosavi, and M. Fakhrahmad, “A semantic approach to extractive multi-document summarization: Applying sentence expansion for tuning of conceptual densities,” Information Processing & Management, vol. 57, no. 6, p. 102341, Nov. 2020, doi: 10.1016/j.ipm.2020.102341.

[8] A. Qaroush, I. Abu Farha, W. Ghanem, M. Washaha, and E. Maali, “An efficient single document Arabic text summarization using a combination of statistical and semantic features,” Journal of King Saud University - Computer and Information Sciences, Mar. 2019, doi: 10.1016/j.jksuci.2019.03.010.

[9] R. Rautray and R. C. Balabantaray, “An evolutionary framework for multi document summarization using Cuckoo search approach: MDSCSA,” Applied Computing and Informatics, vol. 14, no. 2, pp. 134–144, Jul. 2018, doi: 10.1016/j.aci.2017.05.003.

[10] S. Lamsiyah, A. El Mahdaouy, B. Espinasse, and S. El Alaoui Ouatik, “An unsupervised method for extractive multi-document summarization based on centroid approach and sentence embeddings,” Expert Systems with Applications, vol. 167, p. 114152, Apr. 2021, doi: 10.1016/j.eswa.2020.114152.

[11] H. Oliveira et al., “Assessing shallow sentence scoring techniques and combinations for single and multi-document summarization,” Expert Systems with Applications, vol. 65, pp. 68–86, Dec. 2016, doi: 10.1016/j.eswa.2016.08.030.

[12] R. Rautray and R. C. Balabantaray, “Bio-inspired approaches for extractive document summarization: A comparative study,” Karbala International Journal of Modern Science, vol. 3, no. 3, pp. 119–130, Jul. 2017, doi: 10.1016/j.kijoms.2017.06.001.

[13] R. Rautray and R. C. Balabantaray, “Cat swarm optimization based evolutionary framework for multi document summarization,” Physica A: Statistical Mechanics and its Applications, vol. 477, pp. 174–186, Jul. 2017, doi: 10.1016/j.physa.2017.02.056.

[14] J. M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Pérez, “Comparison of automatic methods for reducing the Pareto front to a single solution applied to multi-document text summarization,” Knowledge-Based Systems, vol. 174, pp. 123–136, Jun. 2019, doi: 10.1016/j.knosys.2019.03.002.

[15] E. Linhares Pontes, S. Huet, J.-M. Torres-Moreno, and A. C. Linhares, “Compressive approaches for cross-language multi-document summarization,” Data & Knowledge Engineering, vol. 125, p. 101763, Jan. 2020, doi: 10.1016/j.datak.2019.101763.

[16] K. Yao, L. Zhang, T. Luo, and Y. Wu, “Deep reinforcement learning for extractive document summarization,” Neurocomputing, vol. 284, pp. 52–62, Apr. 2018, doi: 10.1016/j.neucom.2018.01.020.

[17] A. Ghadimi and H. Beigy, “Deep submodular network: An application to multi-document summarization,” Expert Systems with Applications, vol. 152, p. 113392, Aug. 2020, doi: 10.1016/j.eswa.2020.113392.

[18] Y. Wu, Y. Li, and Y. Xu, “Dual pattern-enhanced representations model for query-focused multi-document summarisation,” Knowledge-Based Systems, vol. 163, pp. 736–748, Jan. 2019, doi: 10.1016/j.knosys.2018.09.035.

[19] J. M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Pérez, “Experimental analysis of multiple criteria for extractive multi-document text summarization,” Expert Systems with Applications, vol. 140, p. 112904, Feb. 2020, doi: 10.1016/j.eswa.2019.112904.

[20] L. Marujo et al., “Exploring events and distributed representations of text in multi-document summarization,” Knowledge-Based Systems, vol. 94, pp. 33–42, Feb. 2016, doi: 10.1016/j.knosys.2015.11.005.

[21] M. Rajangam and C. Annamalai, “Extractive document summarization using an adaptive, knowledge based cognitive model,” Cognitive Systems Research, vol. 56, pp. 56–71, Aug. 2019, doi: 10.1016/j.cogsys.2018.11.005.

[22] J. V. Tohalino and D. R. Amancio, “Extractive multi-document summarization using multilayer networks,” Physica A: Statistical Mechanics and its Applications, vol. 503, pp. 526–539, Aug. 2018, doi: 10.1016/j.physa.2018.03.013.

[23] A. John, P. S. Premjith, and M. Wilscy, “Extractive multi-document summarization using population-based multicriteria optimization,” Expert Systems with Applications, vol. 86, pp. 385–397, Nov. 2017, doi: 10.1016/j.eswa.2017.05.075.

[24] T. Uçkan and A. Karcı, “Extractive multi-document text summarization based on graph independent sets,” Egyptian Informatics Journal, vol. 21, no. 3, pp. 145–157, Sep. 2020, doi: 10.1016/j.eij.2019.12.002.

[25] J. M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Pérez, “Extractive multi-document text summarization using a multi-objective artificial bee colony optimization approach,” Knowledge-Based Systems, vol. 159, pp. 1–8, Nov. 2018, doi: 10.1016/j.knosys.2017.11.029.

[26] J. Chen and H. Zhuge, “Extractive summarization of documents with images based on multi-modal RNN,” Future Generation Computer Systems, vol. 99, pp. 186–196, Oct. 2019, doi: 10.1016/j.future.2019.04.045.

[27] J.U. Heu, I. Qasim, and D.-H. Lee, “FoDoSu: Multi-document summarization exploiting semantic analysis based on social Folksonomy,” Information Processing & Management, vol. 51, no. 1, pp. 212–225, Jan. 2015, doi: 10.1016/j.ipm.2014.06.003.

[28] D. Patel, S. Shah, and H. Chhinkaniwala, “Fuzzy logic based multi document summarization with improved sentence scoring and redundancy removal technique,” Expert Systems with Applications, vol. 134, pp. 167–177, Nov. 2019, doi: 10.1016/j.eswa.2019.05.045.

[29] H. Kwon, B.-H. Go, J. Park, W. Lee, Y. Jeong, and J.-H. Lee, “Gated dynamic convolutions with deep layer fusion for abstractive document summarization,” Computer Speech & Language, vol. 66, p. 101159, Mar. 2021, doi: 10.1016/j.csl.2020.101159.

[30] P. Verma and H. Om, “MCRMR: Maximum coverage and relevancy with minimal redundancy based multi-document summarization,” Expert Systems with Applications, vol. 120, pp. 43–56, Apr. 2019, doi: 10.1016/j.eswa.2018.11.022.

[31] A. Qaroush, I. Abu Farha, W. Ghanem, M. Washaha, and E. Maali, “An efficient single document Arabic text summarization using a combination of statistical and semantic features,” Journal of King Saud University - Computer and Information Sciences, vol. 33, no. 6, pp. 677–692, Jul. 2021, doi: 10.1016/j.jksuci.2019.03.010.

[32] Z. Ji, Y. Zhao, Y. Pang, and X. Li, “Cross-modal guidance based auto-encoder for multi-video summarization,” Pattern Recognition Letters, vol. 135, pp. 131–137, Jul. 2020, doi: 10.1016/j.patrec.2020.04.011.

[33] M. Rajangam and C. Annamalai, “Extractive document summarization using an adaptive, knowledge based cognitive model,” Cognitive Systems Research, vol. 56, pp. 56–71, Aug. 2019, doi: 10.1016/j.cogsys.2018.11.005.

[34] D. Wang, H. Fan, and J. Liu, “Learning with joint cross-document information via multi-task learning for named entity recognition,” Information Sciences, vol. 579, pp. 454–467, Nov. 2021, doi: 10.1016/j.ins.2021.08.015.

[35] S. Afantenos, V. Karkaletsis, and P. Stamatopoulos, “Summarization from medical documents: a survey,” Artificial Intelligence in Medicine, vol. 33, no. 2, pp. 157–177, Feb. 2005, doi: 10.1016/j.artmed.2004.07.017.

[36] J. Chen and H. Zhuge, “Summarization of scientific documents by detecting common facts in citations,” Future Generation Computer Systems, vol. 32, pp. 246–252, Mar. 2014, doi: 10.1016/j.future.2013.07.018.

[37] M. Mohd, R. Jan, and M. Shah, “Text document summarization using word embedding,” Expert Systems with Applications, vol. 143, p. 112958, Apr. 2020, doi: 10.1016/j.eswa.2019.112958.

[38] https://duc.nist.gov/data.html.

[39] http://mlg.ucd.ie/datasets/bbc.html.

[40] https://www.ncbi.nlm.nih.gov/research/pubtator-api.