A Novel Summarization-based Approach for Feature Reduction Enhancing Text Classification Accuracy

S. Rahamat Basha
Department of Computer Science & Technology
Sri Krishnadevaraya University
Andhra Pradesh, India

J. Keziya Rani
Department of Computer Science & Technology
Sri Krishnadevaraya University
Andhra Pradesh, India

J. J. C. Prasad Yadav
Department of Computer Science and Engineering, Rajeev Gandhi Memorial College of Engineering and Technology
Andhra Pradesh, India

***Abstract*—Automatic summarization is the process of shortening one (in single document summarization) or multiple documents (in multi-document summarization). In this paper, a new feature selection method for the nearest neighbor classifier by summarizing the original training documents based on sentence importance measure is proposed. Our approach for single document summarization uses two measures for sentence similarity: the frequency of the terms in one sentence and the similarity of that sentence to other sentences. All sentences were ranked accordingly and the sentences with top ranks (with a threshold constraint) were selected for summarization. The summary of every document in the corpus is taken into a new document used for the summarization evaluation process.**

Keywords-summarization; dimension reduction; feature selection; feature extraction; feature clustering; text classification

I. INTRODUCTION

In text classification, the dimensionality of the feature vector is usually huge because the input document set consists of big data and many terms [1, 2]. The major approaches for feature reduction are feature selection [3-10] and feature extraction [11, 12]. Feature extraction approaches are computationally more extensive and more effective than feature selection methods. Another feature reduction technique is feature or attribute clustering, in which similar attributes are grouped into a cluster and each cluster can be considered as an attribute. With clustering noisy data, the repetition of text remains leading to less data transfer rate [13, 14]. To reduce this in pre-processing, the unnecessary words which do not support the classification task (articles, verbs, etc.) are removed. Document summarization is the process that extracts a few sets of sentences and conveys the original meaning of the base corpus. There are two major text summarization approaches: extractive and abstractive. In the extractive approach, sentences or parts of sentences are extracted from the text. In the abstractive method a summary of the text is generated by using some natural language processing techniques. The number of digital documents that are added to the World Wide Web is increasing rapidly, and providing the most relevant information to the user is a big problem. Each digital document consists of many words, but only a few of them are informative. Extracting proper sentence features from documents and categorizing the documents by those features is important and it reflects classification computational efficiency [23-27].

II. RELATED WORK

As shown in Figure 1, the document summarization can be mainly dividedt in the contexts Extractive vs. Abstractive, Generic vs. Query-based, and Single vs. Multiple.

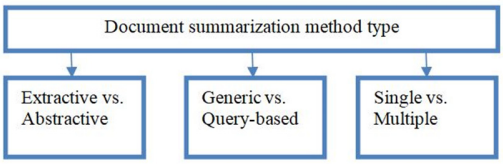

Fig. 1. Document summarization models

A. Extractive vs. Abstractive

In the extractive summarization process, the more important sentences are selected (according to weights or similarities or degrees) from the original document. In abstractive summarization, a summary of the corpus which reflects the main theme of the original document is generated.

B. Generic vs. Query-based

A generic summarization method does not need any queries and focuses on extractive summarization, while in query-based summarization, the summary related to the query is generated.

C. Single vs. Multiple

Single document summarization is the process which summarizes the text of a document whereas multi-documents methods can summarize more than one document at a time. After selecting a suitable feature set, the test document has to be classified by a suitable classification method. Rule-based classification is accurate if the rules are written by experts and easily controlled if the number of rules is small, but if it increases, or the rules conflict with each other, the processing becomes more difficult. If the target domain changes, the rules have to be reconstructed. The machine learning-based approach is domain-independent and gives high predictive performance but training data required [15].

1) Bag of Words Method

Corresponding author: S. Rahamat Basha (basha.ste@gmail.com)

The corpus of documents is generally represented in a bag of words model which is the collection of feature vectors. Let $D = \{d_1, d_2, \ldots\ldots d_n\}$ represent the corpus or document set, where each element of belongs to one class of the set $\{c_1, c_2, \ldots c_p\}$. If a document with only one class label is included in *D*, then it is used for binary or multiclass classification. If a document is included multiple times with multiple class labels in *D*, then it is used for multi-labeled classification.

2) Feature Reduction Method – Example Illustration

Generally, the process of feature reduction can be carried out in two ways, i.e. by feature selection, and feature extraction. The feature selection approach is used to select the subset of original features as the feature set which will be used as input to classification. For example, Information Gain (IG) is better suited to select the most valuable features in many information retrievals. The weight of feature t_k calculated in IG is:

$$w(t_k) = \sum_c \mathrm{P(C)logP(C)} + \mathrm{P}(t_k)\sum_{\mathbf{c}} \mathrm{P(C}|t_k)\mathrm{logP(C}|t_k) + \mathrm{P}(\bar{t}_k)\sum_{\mathrm{c}} \mathrm{P(C}|\bar{t}_k)\mathrm{logP(C}|\bar{t}_k) \quad (1)$$

TABLE I. SAMPLE BAG WITH 10 WORDS OF 10 DOCUMENTS BELONGING TO 2 CLASSES

	office (W_1)	apartment (W_2)	floor (W_3)	line (W_4)	bedroom (W_5)	kitchen (W_6)	building (W_7)	internet (W_8)	WC (W_9)	fridge (W_{10})	class
d_1	1	0	0	1	0	1	0	0	0	1	c_1
d_2	0	0	0	0	0	2	1	1	0	0	c_1
d_3	0	0	0	0	0	0	0	1	0	0	c_1
d_4	0	0	1	2	0	1	1	2	0	1	c_1
d_5	0	0	0	0	1	1	0	0	1	0	c_2
d_6	1	2	1	0	0	1	0	0	1	0	c_2
d_7	2	3	1	0	3	1	1	0	1	0	c_2
d_8	0	1	1	0	1	1	0	0	0	0	c_2
d_9	1	1	1	0	1	0	0	0	0	0	c_2
d_{10}	1	1	1	0	1	0	0	0	0	0	c_2

3) Feature Extraction – Example Illustration

The extracted features in feature extraction approaches are obtained by some processes through algebraic transformation. An Incremental Orthogonal Centroid (IOC) algorithm extracts features as follows [8]: Let some set of documents of a corpus be represented as an $m \times n$ matrix with $X \epsilon\ R^{m \times n}$. Here, *m* stands for the number of documents in the corpus and *n* for the dimensionality of the corpus. The IOC algorithm finds an optimal transformation matrix $F^* \epsilon\ R^{m \times k}$, $k \leq n$, where *k* is the number of features or dimensionality of the projected matrix.

$$F^* = \arg\max\ trace(F^T S_b F)$$

where $F \epsilon R^{m*k}\ and\ F^T F = I$ and

$$S_b = p(C_q)(M_q - M_{all})(M_{q-}M_{all})^T \quad (2)$$

with M_q being the mean vector of class C_q, $P(C_q)$ being the prior probability for a pattern belonging to class C_q and M_{all} being the mean vector of all patterns.

4) Feature Clustering – Modeling Illustration

Generally, clustering means grouping, and in feature clustering, similar features are grouped in one cluster which is treated as a single feature. In hard feature clusters [14, 16-18], the threshold considered for membership in clustering is generally small, which results in a smaller number of features in each cluster. In soft clustering, the threshold size is generally a little larger and it effectively reduces dimensionality, but unimportant features may fall into the clusters. The mixed or medium clustering lies between the above two. The new feature set $W' = \{w'_1, w'_2, \ldots\ldots w'_k\}$ corresponds to partition $\{w_1, w_2 \ldots . w_K\}$ of the original feature set *W*, i.e. $W_t \cap W_q = \Phi$, where $1 \leq q, t \leq k$ and $t \neq q$. Let D be the bag of words consisting of all original documents with m features and D' be the bag of words after projecting D with a new K feature where *K* is less than or equal to the number of features in D.

$$d'_{it} = \sum_{w_j \epsilon w_t} d_{ij} \quad (3)$$

This is the linear sum of the feature values in W_t.

III. PROPOSED SUMMARIZATION METHOD FOR EACH DOUMENT OF EVERY CATEGORY

Our summarization method uses the concept of neighbor with the measures of statistical evaluation. Figure 2 illustrates the framework of the proposed method. This approach consists of four steps.

1. Each pair of sentence similarity is calculated.
2. The term frequency (TF) of each sentence is calculated.
3. The score for each sentence of a document is computed concerning all documents that belong to the same category.
4. Ranks to each sentence of that document are assigned according to their computed scores.

Let d_0 be the document input to be summarized. Our system finds the documents near d_0 from the same category collection. The nearest neighbor documents that are similar to the specified document are identified by using the nearest neighbor search technique. The standard cosine measure is used to find the similarity between two documents. Let d_i, d_j be two documents. The similarity between them is calculated as follows:

$$sim_{doc}(d_i, d_j) = \frac{\vec{d_i} \cdot \vec{d_j}}{\left\|\vec{d_i}\right\| \times \left\|\vec{d_j}\right\|} \quad (4)$$

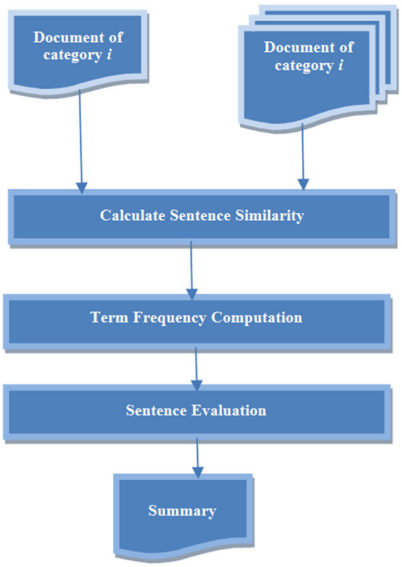

Fig. 2. The proposed single document summarization architecture

Affinity Graph Construction method is used to show the relationship between sentences. Let $G = (S, E)$ be an undirected graph where S represents a set of sentences, E represents a set of edges, and all edges among sentences are associated with a weight. The edge between sentence i and sentence j is represented as e_{ij}, and the corresponding weight of that edge is the similarity value of those sentences. The edges of the graph G are divided into two categories as follows: D_1 is a document with $D_1S_1, D_1S_2, D_1S_3, D_1S_4$, and D_2 a document with $D_2S_1, D_2S_2, D_2S_3, D_2S_4$.

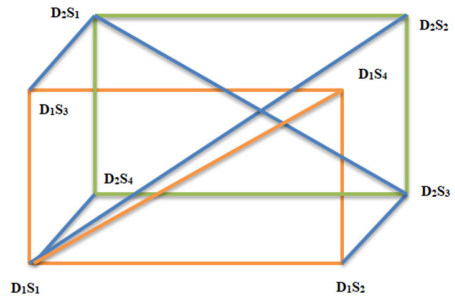

Fig. 3. Example diagram showing with-in document and cross-document sentence similarities

A. With-in Document Link (WDL) and Cross Document Link (CDL)

The edge between D_1S_1 and D_1S_2 is an example of WDL while the edge between D_1S_3 and D_2S_1 is an example of cross-document links. Y is the similarity between statements D_1S_1 & D_1S_2. The link between two sentences from the same document is a WDL and provides local information. The link between two sentences which belong to two different documents is a CDL and provides global information. Both these link kinds are incorporated in a global affinity graph G and are represented by an adjacency matrix M in which each entry is the weight of the sentences corresponding to the matrix row and column. The matrix is defined as follows:

$$M_{ij} = \begin{cases} \lambda \times sim_{sen}(s_i, s_j), & if\, i \neq j \\ 0, & otherwise \end{cases} \quad (5)$$

The confidence value is denoted by λ, it is 1 if the link is WDL and $sim_{doc}(d_k d_1)$ if it is otherwise. The matrix M is normalized as follows:

$$\tilde{M}_{i,j} = \begin{cases} M_{i,j} / \sum_{j=1}^{|s|} M_{i,j}, & if \sum_{j=1}^{|s|} M_{i,j} \neq 0 \\ 0, & otherwise \end{cases} \quad (6)$$

B. Term Frequency Computation

It is important to calculate the importance of the sentences in summarization. The sentences with importance contain words that occur frequently. Term frequency is noting how many times a term is present in a document. The similarity measure of each sentence with all the sentences in the document illustrates the importance of that sentence in the knowledge base, which is the collection of all the documents belonging to a category. Term frequency is a good sentence extraction method because it gives the priority of the sentence in a document while it can be computed very easily. In a sentence, the TF of each term is calculated separately and the sum of all these term frequencies gives $T(S_i)$), for the sentence (S_i) is defined as follows:

$$T(s_i) = \sum_{i=0} tf(t_i) \quad (7)$$

where $tf(t_i)$ is the term frequency of term t_i, the i^{th} term in the sentence.

C. Sentence Evaluation

We used the Global Affinity Graph G to calculate the scores for the sentences which are linked to it and denoted as $SSCore(s_i)$ where sS_i represents sentence i as follows:

$$SSCore(s_i) = \sum_{alli \neq j} T_{i,j} \tilde{M}_{i,j}$$

where:

$$T_{i,j} = \begin{cases} \frac{T(s_i)}{|T(s_i) - T(s_j)|}, & if |T(s_i) - T(s_j)| \neq 0 \\ 1, & otherwise \end{cases} \quad (8)$$

D. Document Level Redundancy Removing

The sentences which are highly overlapped with others are removed by applying a penalty on them to avoid redundant information in the summary. The redundancy removing algorithm is [3]:

1. Initialize:

$$\text{Set } A = \varphi,$$
$$\text{Set } B = \{S_i\} where\ i = 1,2,3, \ldots . p, \text{ and}$$
$$ARScore(S_i) = SScore(S_i), i = 1,2, \ldots . p$$

2. The sentences in set B are to be sorted according to their affinity scores (i.e. in descending order).
3. The sentence S_i with high score in set B is to move to set A.

4. A penalty on the affinity rank scores is applied to the sentences that are in set B with respect to the similarity of the sentence that is recently moved to set A.
5. From step 2 repeat the process until set $B = \varphi$.

The algorithm outputs the final score for each sentence in the specified document and based on these scores the sentences were ranked. Depending on the summary limit specified, the highest-ranked sentences were selected. We selected 5 and 10 lines from each document which are summarized and labeled as in the corpus in TCEJ (Text Categorization Environment-implemented in Java) experiment both times [4]. We got 150 labeled summarized training documents and tested them with the reut2-003 test document. We assumed that the results would be better for Mutual Information (MI), than for Information Gain (IG) because each document is summarized for all documents belonging to that category. We got similar results for the MI, IG, DF feature selection techniques for the summarized training corpus with the results obtained with the original training corpus.

IV. RESULTS AND ANALYSIS

Two document collections were used for the TCEJ experiment [4]. A self-made data set was used to train the classifiers and Reuters 21578 corpus was used as the test data set [19-22]. Generally, the training corpus should be very suitable for every category and should be small to reduce the needed running time. This self-made corpus was collected from standard categories Business, Science, Sports, Health, Education, Movies, Travel and consisted of 150 documents.

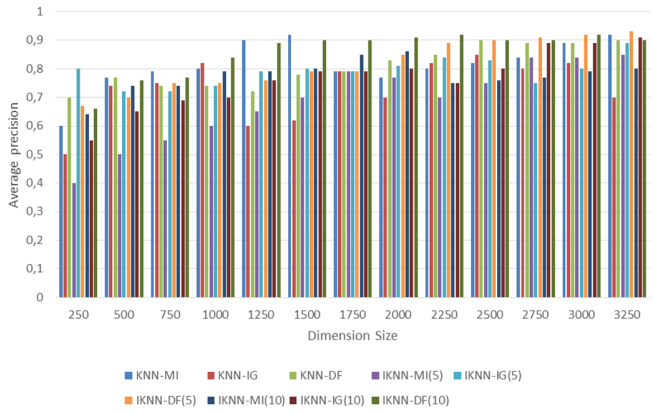

Fig. 4. Average precision comparison of MI, IG, and DF techniques

TABLE II. SELF-MADE DOCUMENT COLLECTION AND CATEGORIES

Category	# of documents
Sports	30
Health	30
Science	27
Business	23
Education	24
Travel	06
Movies	10

TABLE III. RESULT COMPARISON OF FEATURE VECTOR FOR NORMAL TRAINING DOCUMENT SET

Corpus	Stop words	Stemming	# of processed terms	# of unique terms
Self-made	No	No	105443	12819
Self-made	Yes	No	105443	12660
Self-made	No	Porter	105443	8878
Self-made	Yes	Porter	105443	8697
Self-made	No	Lancaster	105443	7490
Self-made	Yes	Lancaster	105443	7288

TABLE IV. RESULT COMPARISON OF FEATURE VECTOR FOR SUMMARIZED TRAINING DOCUMENT SET *

Corpus	Stop words	stemming	# of processed terms	# of unique terms
Self-made	No	No	19332	5150
Self-made	Yes	No	19332	5025
Self-made	No	Porter	19332	3977
Self-made	Yes	Porter	19332	3838
Self-made	No	Lancaster	19332	3510
Self-made	Yes	Lancaster	19332	3353

* each document consists of 5 summarized sentences

TABLE V. RESULTS COMPARISON OF FEATURE VECTOR FOR SUMMARIZED TRAINING DOCUMENT SET *

Corpus	Stop words	stemming	# of processed terms	# of unique terms
Self-made	No	No	37323	7497
Self-made	Yes	No	37323	7358
Self-made	No	Porter	37323	5479
Self-made	Yes	Porter	37323	5330
Self-made	No	Lancaster	37323	4734
Self-made	Yes	Lancaster	37323	4563

*each document consists of 10 summarized sentences

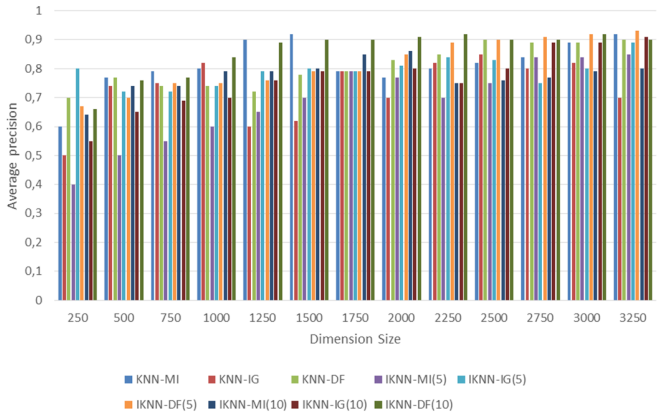

Fig. 5. Micro and Macro averaged accuracy of MI technique

V. CONCLUSION AND FUTURE WORK

Our proposed single document summarization method by using all documents belonging to a category consists of two statistical measures, term frequency, and sentence similarity. We can conclude that the summary generated in this work is more efficient. The feature selection techniques MI, IG, DF gave good results. MI and IG gave similar results for binary problems. Regarding future work, we aim at implementing multi-document summarization, in which all documents belonging to one category are summarized at a time and one

single summarized document is constructed for each category. Another aim is to implement some Hidden Markov Model methods for the unigram, Bi-gram, N-gram features that are selected from the summarized documents which were generated by single document summarization and multi-document summarization.

REFERENCES

[1] J. Y. Jiang, R. J. Liou, S. J. Lee, "A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification", IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No. 3, pp. 335–349, 2011

[2] D. D. Lewis, "Feature selection and feature extraction for text categorization", Workshop on Speech and Natural Language, Harriman, USA, February 23-26, 1992

[3] X. Wan, J. Xiao, "Exploiting neighborhood knowledge for single document summarization and keyphrase extraction", ACM Transactions on Information Systems,Vol. 28, No. 2, Article 8, 2010

[4] M. Niepert, "An experiment system for textcClassification", available at: https://www.semanticscholar.org/paper/An-Experiment-System-for-Text-Classification-Niepert/32be395201b132eb64939a1ca8541efd0f1e8984, 2005

[5] H. Kim, P. Howland, H. Park, "Dimension reduction in text classification with support vector machines", Journal of Machine Learning Research, Vol. 6, pp.37-53, 2005

[6] A. L. Blum, P. Langley, "Selection of relevant features and examples in machine learning", Artificial Intelligence, Vol. 97, pp. 245-271, 1997

[7] E. F. Combarrow, E. Montanes, I. Diaz, J. Ranilla, R. Mones, "Introducing a family of linear measures for feature selection in text categorization", IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 9, pp. 1223-1232, 2005

[8] Y. Jan, B. Zhang, N. Liu, S. Yan, Q. Cheng, W. Fan, Q. Yang, W. Xi, Z. Chen, "Effective and efficient dimensionality reduction for large-scale and streaming data preprocessing", IEEE Transactions on Knowledge and Data Engineering, Vol. 18, No. 3, pp. 320-333, 2006

[9] M. Alghobiri, "A comparative analysis of classification algorithms on diverse datasets", Engineering, Technology & Applied Science Research, Vol. 8, No. 2, pp. 2790-2795, 2018

[10] E. Jamalian, R. Foukerdi, "A hybrid data mining method for customer churn prediction", Engineering, Technology & Applied Science Research, Vol. 8, No. 3, pp. 2991-2997, 2018

[11] D. Koller, M. Sahami, "Toward optimal feature selection", 13th International Conference on Machine Learning, Bari, Italy, July 3-6, 1996

[12] R. Kohavi, G. H. John, "Wrappers for feature subset selection", Artificial Intelligence, Vol. 97, No. 1-2, pp. 273-324, 1997

[13] Y. Yang, J. O. Pederson, "A comparative study on feature selection in text categorization", 14th International Conference on Machine Learning, San Francisco, USA, July 8-12, 1997

[14] N. Slonim, N. Tishby, "The power of word clusters for text classification", 23rd European Colloquium on Information Retrieval Research, 2001

[15] Y. Sasaki, Automatic Text Classification, Lecture notes, University of Manchester, available at: http://www.nactem.ac.uk/dtc/DTC-Sasaki.pdf, 2008

[16] L. D. Baker, A. McCallum, "Distributional clustering of words for text classification", 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, August 24-28, 1998

[17] R. Bekkerman, R. EI-Yaniv, N. Tishhby, Y. Winter, "Distributional word clusters versus words for text categorization", Journal of Machine Learning Research, Vol. 3, pp. 1183-120, 2003

[18] I. S. Dhilllon, S. Mallela, R. Kumar, "A divisive information theoretic feature clustering algorithm for text classification", Journal of Machine Learning Research, Vol. 3, pp. 1265-1287, 2003

[19] https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection

[20] https://martin-thoma.com/nlp-reuters/

[21] http://www.daviddlewis.com/resources/testcollections/reuters21578/

[22] http://disi.unitn.it/moschitti/corpora.htm

[23] S. Rahamat Basha, J. Keziya Rani, J. J. C. Prasad Yadav, G. Ravi Kumar, "Impact of feature selection techniques in Text Classification: an experimental study", J. Mech. Cont.& Math. Sci., Special Issue, No. 3, pp. 39-51, 2019

[24] G. Ravi Kumar, K. Nagamani, "A framework of dimensionality reduction utilizing PCA for neural network prediction", International Conference on Data Science and Management, Bhubaneswar, USA, February 22-23

[25] G. Ravi Kumar, K. Nagamani, "Banknote authentication system utilizing deep neural network with PCA and LDA machine learning techniques", International Journal of Recent Scientific Research, Vol. 9, No. 12, pp. 30036-30038, 2018

[26] M. V. Lakshmaiah, G. Ravi Kumar, G. Pakardin, "Framework for finding association rules in big data by using Hadoop Map/Reduce tool", International Journal of Advance and Innovative Research, Vol. 2, No. 1(I), pp. 6-9, 2015

[27] G. Ravi Kumar, G. A. Ramachandra, K. Nagamani, "An efficient prediction of breast cancer data using data mining techniques", International Journal of Innovations in Engineering and Technology, Vol. 2, No. 4, pp. 139-144, 2013