ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

Classification of Text Documents: Compressed Dictionary Based Approach

Vidya Lakshmi[1], Sachin Bharadwaj[2] , Prashal I S[2], Bharath Bhushan S. N.[2]

Department of Computer Applications, Sahyadri College of Engineering & Management, Mangalore, Karnataka, India

ABSTRACT- Internet is a pool of information, which contains billions of text documents which are stored in compressed format. In literature we can find many text classification algorithms which work on uncompressed text documents. In this paper, we propose a novel representation scheme for a given text document using compression technique. Further, centroid classifier is also designed for classification of text documents. For the purpose of compression, LZW compression technique is used and the compressed dictionary representation obtained by LZW technique is used as representative for the text document. Extensive experimentation is carried out on seven datasets, out of which three are our own datasets and remaining four are publically available datasets resulting with approximately 89% of F-measure.

KEYWORDS: Text classification, Text compression, LZW compression technique.

I. INTRODUCTION

Internet is the rapidly growing information gallery that contains rich textual information. This rapid growth makes it difficult for the users to locate relevant information quickly on the web. Document retrieval, categorization, routing and filtering systems are often based on text classification. Text classification problem can be stated as follows: given a set of labeled examples belonging to two or more classes, we classify a new test document to a class with the highest similarity. Text classification presents many challenges and difficulties. Firstly, it is difficult to capture high-level semantics and abstract concepts of natural languages just from a few key words and the same word can represent different meanings. Secondly, it is difficult to handle high dimensionality and variable lengths of text documents.

Text Documents are the most common type of information store house especially with the increased use of the internet. Internet web pages, e-mails, e-news feeds newsgroup messages have millions or billions of text documents. The web pages that are available in the internet are stored in the compressed format. Data mining activities such as document classification and clustering are carried out these data by decompressing the data and taking it back to the standard format. These processes of decompressing and performing mining activities consume more computational time. However to the best of our knowledge, nowhere in the literature we can find any works on classification of text documents in text compressed format. This motivated us to take up this work for design of text classification using text compression representation as a new representation method.

The rest of the paper is organized as follows. In section 2 a brief literature survey on the text classification is presented. In section 3 we present the model based on LZW compression technique. In section 4 we discuss about experimentation details and comparative analysis. In section 5 we present conclusion along with future work.

II. RELATED WORK

In automatic text classification, it has been proved that the term is the best unit for text representation and classification [1]. Though a text document expresses vast range of information, unfortunately, it lacks the imposed structure of traditional database. Therefore, unstructured data, particularly free running text data has to be transformed into a structured data. To do this, many pre-processing techniques are proposed in literature [2, 3]. After converting an unstructured data into a structured data, we need to have an effective document representation model to build an efficient classification system. Bag of Word (BoW) is one of the basic methods of representing a document. The BoW is used to form a vector representing a document using the frequency count of each term in the

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

document. This method of document representation is called as a Vector Space Model (VSM) [4]. The major limitation of VSM is that the correlation and context of each term is lost which is very important in understanding a document. Li and Jain [5] used binary representation for given document. The major drawback of this model is that it results in a huge sparse matrix, which raises a problem of high dimensionality. Another approach [6] uses multi-word terms as vector components to represent a document. But this method requires a sophisticated automatic term extraction algorithms to extract the terms automatically from a document. Wei et al., (2008) proposed an approach called Latent Semantic Indexing (LSI) [7] which preserves the representative features for a document. The LSI preserves the most representative features rather than discriminating features. Thus to overcome this problem, Locality Preserving Indexing (LPI) [8] was proposed for document representation. The LPI discovers the local semantic structure of a document. Unfortunately LPI is not efficient in time and memory [9]. Choudhary and Bhattacharyya (2002) [10] used Universal Networking Language (UNL) to represent a document. The UNL represents the document in the form of a graph with words as nodes and relation between them as links. This method requires the construction of a graph for every document and hence it is unwieldy to use for an application where large numbers of documents are present.

After giving an effective representation for a document, the task of text classification is to classify the documents to the predefined categories. In order to do so, many statistical and computational models have been developed based on Naïve Bayes classifier [11], K-NN classifier [12], Centroid Classifier [13], Decision Trees [14], Rocchio classifier [15], Support Vector Machines [16].

Although many text document representation models are available in literature, frequency-based BoW model gives effective results in text classification task. Indeed, till date the best multi-class, multi-labelled categorization results for well known datasets are based on BoW representation [17]. Unfortunately, BoW representation scheme has its own limitations. Some of them are: high dimensionality of the representation, loss of correlation with adjacent words and loss of semantic relationship that exist among the terms in a document [18]. Also the main problem with the frequency based approach is that given a term, with lesser frequency of occurrence may be appropriate in describing a document, whereas, a term with the higher frequency may have a less importance. Unfortunately, frequency-based BoW methods do not take this into account [10].

All the mentioned algorithms works on uncompressed documents. Whereas the challenging and required is to classify documents at compression level. In literature we can find many compression techniques which are used for the effective representation of data in compressed format. In this paper we consider only the lossless compression schemes. Run Length Encoding (RLE) [19] is a simple and popular data compression algorithm. It is based on the idea to replace a long sequence of the same symbol by a shorter sequence. Huffman compression [20] it is a statistical lossless compression method that converts characters into variable length bit strings. Huffman compression technique works on frequency of individual symbol. The Huffman algorithm is a so-called "greedy" approach to solving this problem in the sense that at each step, the algorithm chooses the best available option. Lempel–Ziv–Welch (LZW) is a universal lossless data compression algorithm created by Abraham Lempel, Jacob Ziv, and Terry Welch. The LZW compression algorithm organized around a translation table or string table, that maps input characters into the fixed length codes [21]. Among different compression techniques, we have used LZW compression technique. LZW compression is used as the foremost technique, mainly because of its versatility and simplicity. Typically, the LZW compression can compress executable code, text, and similar data files to almost one-half of their original size. It usually uses single codes to replace strings of characters, thereby compressing the data. LZW also gives a good performance when extremely redundant data files are presented to it like computer source code, tabulated numbers and acquired signals. The common compression ratio for these cases is almost in the range of 5:1.Though RLE and Huffman compression techniques are also very simple; they are not suitable for text documents and also these two methods does not provide good compression ratio like LZW method.

III. PROPOSED METHOD

In this paper we are proposing a novel method used for classification of compressed text documents. Normally text documents are available in several formats such as html, xhtml, pdf, plain text etc. The first step is to pre-process the text document, hence to bring them to a common format before processing the text. In the literature we can find many

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

techniques for pre-processing for text documents. They are stop word elimination, stemming, pruning etc as pre-processing steps. In this work we have used only stop word elimination technique. Once the pre-processing is done on training data, the text documents are compressed using LZW compression scheme and a compressed training document library is created. The working principle of LZW compression technique is given as follows.

LZW is a universal lossless compression algorithm which is organized around string table. String table contains strings that have been encountered previously in the text being compressed. It consists of a running sample of strings in the text, so the available strings reflect the statistics of the text. It uses greedy parsing algorithm, where the input string is examined character-serially on one pass, and the longest recognized input string is parsed off each time. A recognized string is one that exists in the string table. Each such added string is assigned a uniquely identified by code value.
The proposed model is of two stages, in which stage one is of creation of knowledgebase in which all pre-processed text data are compressed and preserved, stage two is classification stage in which given unknown sample is classified into its corresponding class label using compression technique.

Algorithm: LZW text compression.
Input: Pool of text data
Output: Pool of compressed text data, String table.
Method:

1. Initialize table to contain single character strings.
2. Prefix string ω ← Read first input character.
3. K← Read next input character
   If no such K (input exhausted) : code (ω) – output; EXIT
4. If ωK exists in string table : ωK- ω; repeat 3;
5. else ωK not in string table : code (ω) – output;
6. ωK – string table;
7. K – ω; repeat Step.

Algorithm end.

At each execution of the basic step an acceptable input string ω has been parsed off. The next character K is read and the extended string ωK is tested to see if it exists in the string table. For each training document we obtain a string table which is referred as dictionary representation and stored in the library. Further, given a test document we obtain dictionary representation and during classification we use string matching based on centroid classification technique. We classify the test document and class label is assigned. The block diagram of the proposed model is as shown in fig 1.

IV. EXPERIMENTATION

In this section, we present the details of the experiments conducted to represent the effectiveness of the proposed method on seven datasets. We have created three datasets of our own and four publically available datasets to evaluate the performance of the proposed model. First dataset consists of three classes and each class consists of five documents. Second dataset consists of five classes and each class consists of ten documents. Third dataset consist of 1000 documents from 10 different classes. Fourth dataset is Google news group dataset which contains one thousand documents from ten different classes and fifth dataset is vehicles Wikipedia [22] used to evaluate a prototype system used for the evaluation of classification performance. Seventh dataset is 20 mini newsgroup dataset. Sixth and seventh datasets are the 20 newsgroups mini and 20 newsgroup large dataset which are publically available dataset consisting collection of 2000 and 20000 newsgroup documents, partitioned evenly across 20 different classes. The first three datasets consists of documents which do not have overlap compared to other publically available datasets. This is considered to study the performance of the proposed model in case of less overlap and large overlap.

We have conducted two sets of experiments; where each set contain three different trails. In first set of experiments, we have used 40% of the database for training and remaining 60 % is used for testing. In second set of experiments, we have used 60 % training and 40 % for testing. Each set of experiments contain three different trials. In each trail we have selected training and testing document randomly. For the purpose of evaluation of results, we have calculated precision, recall and f-measure for each trail. The details of the experiments are shown in the following table 1.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

In table 2, we have listed, max, mean and standard deviation of the results of each data set and it can be seen that mean is comparatively high and standard deviation is very less. This indicates that the proposed method works well even in case of different set of training and testing sets. Also, Tab 2 indicates that, the proposed model performs equally well in case of both large overlap and less overlap cases. The quantitative evaluation of the proposed method is carried out with existing different methods of text classifiers. The proposed method with different type classification techniques are analysed in qualitative comparative analysis is also presented. Table 3 reveals that, all the existing works in the literature are done on the uncompressed text documents. But the proposed model classifies the documents in compressed format also.

V. CONCLUSION

Novel LZW compression based technique for classification of text documents is presented. The proposed method uses LZW compressed dictionary representation scheme for representation of text documents. Using string matching and centroid classification method we have proposed text classification technique. To check the efficiency and the robustness of the proposed models, an extensive experiment is carried out on all the seven dataset. The performance evaluation of the proposed method is carried out by performance measures such as precision, recall and f-measure. Even though, the results are not better than other uncompressed based techniques, they are comparatively equal to them, i.e., approximately 89% of classification accuracy. In this paper we have put forward a new representation model for text classification using compression technique, which is first of its kind. Further, we explore novel proximity measures for comparing text in compressed format which may improve the classification accuracy.

REFERENCES

[1] Rigutini L., 2004. Automatic text processing: Machine learning techniques. Ph.D. Thesis, University of Siena.
[2] Porter M. F., 1980. An algorithm for suffix stripping. Program, vol. 14, no. 3, pp. 130 –137.
[3] Hotho A., A. Nurnberger and G. Paab, 2005. A brief survey of text mining. Journal for Computational Linguistics and Language Technology, vol. 20, pp. 19 – 62.
[4] Salton G., A. Wang and C. S. Yang, 1975. A vector space model for automatic indexing. Communications of the ACM, vol. 18, no. 11, pp. 613 – 620.
[5] Li Y. H and A. K. Jain, 1998. Classification of text documents. The Computer Journal, vol. 41, no. 8, pp. 537 – 546.
[6] Shafiei M, W. Singer, R. Zhang, E. Milios, T. Bin, J. Tougas and R. Spiteri., 2007. Document Representation and Dimension Reduction for Text Clustering. Proceedings of the IEEE 23rd International Conference on Data Engineering, USA, pp. 770 – 779.
[7] Wei, C.P., Yang, C.C., Lin, C.M. (2008) Journal of Decision Support System. 606—620
[8] He X., D. Cai., H. Liu and W. Y. Ma, 2004 (a). Locality preserving indexing for document representation. Proceedings of the International Conference on Research and Development in Information Retrieval, UK, pp. 96 – 103.
[9] Cai D., X. He and J. Han, 2005. Document clustering using locality preserving indexing. IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 12, pp. 1624 –1637.
[10] Choudhary B and P. Bhattacharyya, 2002. Text clustering using universal networking language representation. Proceedings of the 11th International Conference on World Wide Web, USA (http://www-clips.imag.fr/geta/User/wang-ju.tsai/articles/BhChPBh-UNL01.PDF).
[11] Mccallum A. K and Nigam K., 1998. Employing EM in pool-based active learning for text classification. Proceedings of the 15th International Conference on Machine Learning, USA, pp. 350 – 358.
[12] Tan S., 2007. An effective refinement strategy for k-NN text classifier. Journal of Expert Systems with Applications, vol. 30, no. 2, pp. 290 – 298.
[13] Tan S., 2008. An improved centroid classifier for text categorization. Journal of Expert System with Applications, vol. 35, no. 2, pp. 279 – 285.
[14] Wang J., Y. Yao and Z. J. Liu, 2007. A new text classification method based on HMMSVM. Proceedings of the 7th International Symposium on Communications and Information Technologies, Australia, pp. 1516 – 1519.
[15] Lewis D. D., 1992. An evaluation of phrasal and clustered representations on a text categorization task. Proceedings of the 15th Annual International Conference on Research and Development in Information Retrieval, Denmark, pp. 37 – 50.
[16] Mitra V., C. J. Wang and S. Banerjee, 2007. Text classification: A least square support vector machine approach. Journal of Applied Soft Computing, vol. 7, no. 3, pp. 908– 914.
[17] Bekkerman R and J. Allan, 2003. Using bigrams in text categorization. CIIR Technical Report, University of Massachusetts.
[18] Bernotas M., K. Karklius., R. Laurutis and A. Slotkiene, 2007. The peculiarities of the text document representation, using ontology and tagging-based clustering technique. Journal of Information Technology and Control, vol. 36, no. 2, pp. 217 – 220.
[19] P.A. Franaszek (1972), U.S. Patent 3,689,899.
[20] D. A. Huffman., 1952. A method for construction of minimum-redundancy codes. Proceeding of the Institute of Electrical and Radio Engineers, 40(9) pp. 1090 – 1101.
[21] Ziv. J and Lempel A., 1978. Compression of Individual Sequences via Variable-Rate Coding, IEEE Transactions on Information Theory 24 (5), pp. 530–536.
[22] Isa D., L. H. Lee., V. P. Kallimani and R. Rajkumar., 2008. Text document preprocessing with the bayes formula for classification using the support vector machine. IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 9, pp. 23 – 31.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

[23] Xue X. B and Z. H. Zhou., 2009. Distributional features for text categorization. IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 3, pp. 428 – 442.
[24] D.S. Guru, B.S. Harish, and S. Manjunath, "Symbolic Representation of Text Documents" Proceedings of the International ACM Conference, Compute 2010, Bangalore, India.
[25] Dinesh, R., Harish, B. S., Guru, D.S., and Manjunath, S. 2009. Concept of Status Matrix in Text Classification. In the Proceedings of Indian International Conference on Artificial Intelligence, Tumkur, India, pp. 2071 – 2079.

Table 1 : Classification result table on different dataset using proposed model

Dataset	Trails	40 : 60				60:40			
		Precision	Recall	F Measure	Avg F-Measure	Precision	Recall	F Measure	Avg F-Measure
DATASET 1	1	0.8055	0.7777	0.775	0.768	0.8888	0.8333	0.8222	0.767
	2	0.8333	0.7777	0.74		0.8888	0.8333	0.8222	
	3	0.8333	0.7777	0.79		0.7222	0.6666	0.6555	
DATASET 2	1	0.9	0.8666	0.863	0.8	0.9333	0.9	0.8933	0.822
	2	0.8333	0.8	0.796		0.8666	0.8	0.7866	
	3	0.7833	0.7333	0.74		0.8666	0.8	0.7866	
DATASET 3	1	0.7743	0.7983	0.7854	0.774	0.7953	0.7976	0.7959	0.786
	2	0.7882	0.7717	0.7773		0.7762	0.7925	0.7837	
	3	0.7576	0.7633	0.7598		0.7667	0.7925	0.7777	
DATASET 4	1	0.7876	0.7767	0.7819	0.783	0.7901	0.7775	0.7831	0.782
	2	0.7717	0.7333	0.7801		0.7714	0.79	0.78	
	3	0.7929	0.7929	0.7866		0.7941	0.775	0.7839	
DATASET 5	1	0.7983	0.8	0.7984	0.798	0.7803	0.8	0.7894	0.784
	2	0.8089	0.7883	0.7975		0.7714	0.79	0.78	
	3	0.7889	0.81	0.7978		0.7763	0.7925	0.7834	
DATASET 6	1	0.7754	0.7955	0.7828	0.784	0.7944	0.7898	0.7888	0.788
	2	0.787	0.7689	0.775		0.7762	0.7841	0.7799	
	3	0.7994	0.7841	0.7928		0.8044	0.7898	0.7954	
DATASET 7	1	0.7876	0.7758	0.7797	0.776	0.7985	0.77	0.7829	0.776
	2	0.7526	0.7816	0.7473		0.7863	0.7725	0.7787	
	3	0.8124	0.7925	0.8		0.7722	0.76	0.7649	

Table 2 : Max, Mean and Standard deviation of F-Measure

	40 : 60			60 : 40		
	Max	Mean	Standard deviation	Max	Mean	Standard deviation
D1	0.79	0.775	0.025658	0.8222	0.8222	0.096244
D2	0.863	0.796	0.061582	0.8933	0.7866	0.061603
D3	0.7854	0.7773	0.013084	0.7959	0.7837	0.009274
D4	0.7866	0.7819	0.003356	0.7839	0.7831	0.00206
D5	0.7984	0.7978	0.000458	0.7894	0.7834	0.00476

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

D6	0.7928	0.7828	0.008923	0.7954	0.7888	0.007778
D7	0.8	0.7797	0.026581	0.7829	0.7787	0.009417

Table 3 : Quantitative Evaluation Table

Authors	Dataset	Compressed /Uncompressed	Representation Scheme	Classifiers Used	Min F-Measure	Max F – Measure
Xue and Zhou, 2009	Reuters 21578	Uncompressed	Distributional Words	K-NN	0.4950	0.8440
	20Newsgroup	Uncompressed	Distributional Words	AVM	0.8860	0.8870
Guru et al 2010	Vehicles Wikipedia	Uncompressed	Symbolic Representation	SVM	0.8850	0.9050
	20 Mini Newsgroup	Uncompressed			0.8650	0.8950
	Google Newsgroup	Uncompressed			0.8750	0.8900
	Research Article Abstracts	Uncompressed			0.8620	0.9070
Harish et al 2010	Google Newsgroup	Uncompressed	Symbolic Representation		0.8440	0.9220
	Research Article Abstracts	Uncompressed	Symbolic Representation		0.8460	0.9750
Dinesh et al 2009	Vehicles Wikipedia	Uncompressed	Status Matrix	Voting Classifier	0.8850	0.9050
	20 Mini Newsgroup	Uncompressed	Status Matrix	Voting Classifier	0.8650	0.8950
	Google Newsgroup	Uncompressed	Status Matrix	Voting Classifier	0.8800	0.9200
	Research article Abstracts	Uncompressed	Status Matrix	Voting Classifier	0.9180	0.9260
Proposed Method	Dataset 1	**Compressed**	Dictionary Representation	NN Classifier	0.6977	0.8061
	Dataset 2				0.7400	0.8933
	Dataset 3				0.7598	0.7959
	Dataset 4				0.7800	0.7866
	Dataset 5				0.7975	0.7834
	Dataset 6				0.7750	0.7829
	Dataset 7				0.7473	0.8000

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

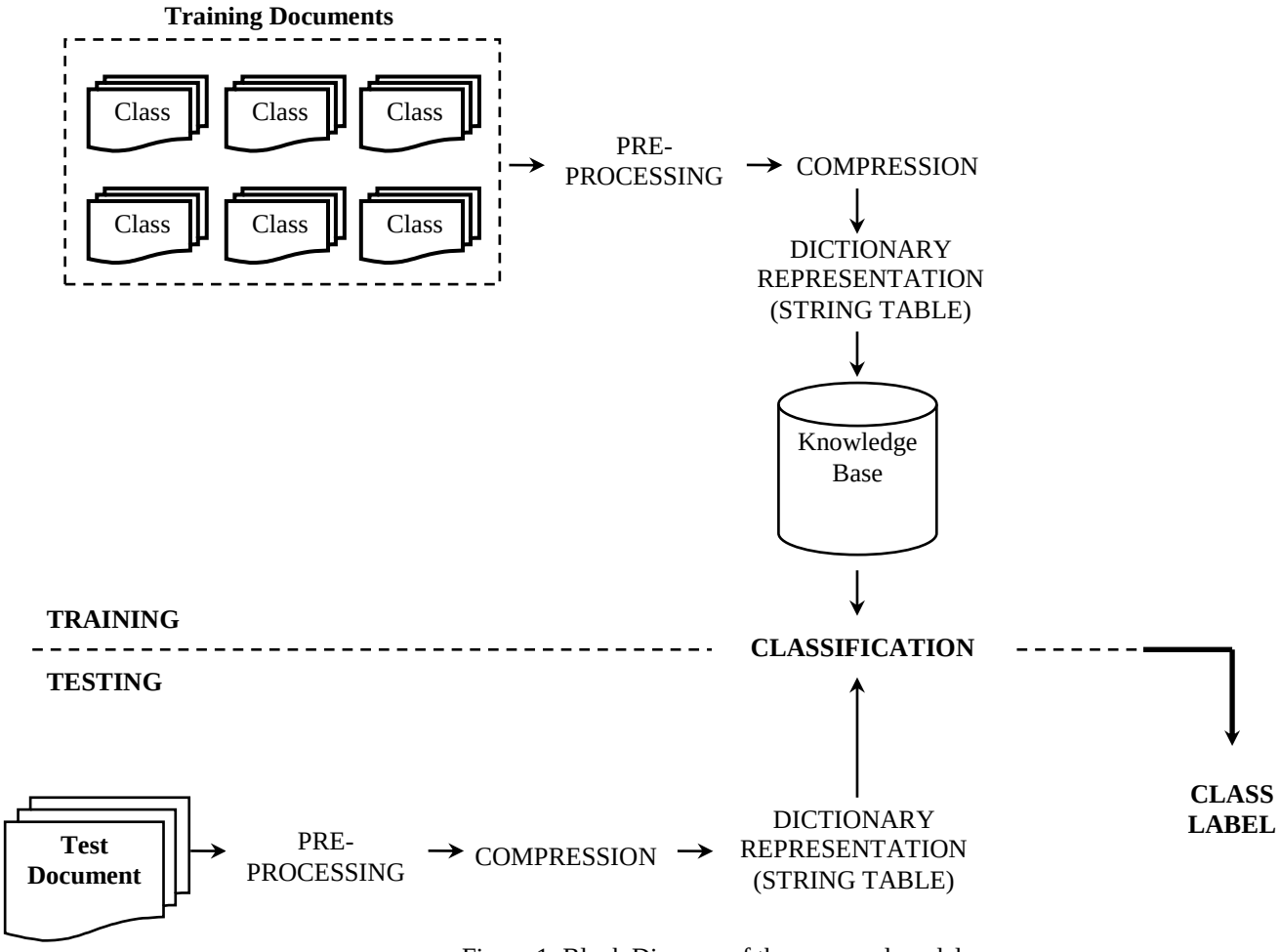

Figure 1: Block Diagram of the proposed model.