ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2020.0315

IMPROVED FEATURE SET EXTRACTION FROM DOCUMENTS USING MODIFIED BAG OF WORDS

R. Sathish Babu and R. Nagarajan
Department of Computer and Information Science, Annamalai University, India

Abstract

In conventional literatures, there are several different methods of collection and extraction and are also used to minimize dimensionality. Traditional methods are intuitively designed to delete redundant and outdated information to help define new test cases more effectively. But the number of specific words in the Bag of Words (BoW) model must be manually calculated, requiring time and work and portability of deficiencies. In addition, the number of codebook vectors in BoW rises as cancer types grow and the efficiency and accuracy of detection are reduced. The BoW model is therefore not ideal for multi-operative failure diagnosis. Therefore, we propose an improved BoW in this paper which selects the number of special terms required to collect cancer diagnostic functions from different documents. The overall recognition and accuracy rates are higher than other existing extraction models. The improved BoW method has been verified to be highly effective in operating conditions that meet the requirements in real time.

Keywords:

Bag of Words, Cancer Document Retrieval, Codebook, Dimensionality Reduction

1. INTRODUCTION

In recent decades, there has been a steady evolution in cancer research [1]. Scientists have applied various methods, such as early stage screening, to identify types of cancer before symptoms occur. In addition, new strategies have been developed for early prediction of the results of cancer treatment. With the advent of new technology in the medical field, large quantities of cancer data have been collected and available to the medical research community. However, predicting a disease result accurately is one of the most important and challenging tasks for doctors. As a result, ML methods have become a popular instrument for medical scientists. These techniques can find and identify patterns and relationships amongst them from complex datasets and can effectively predict future outcomes of a type of cancer.

Fast, reliable text analysis and the extraction of information from free form natural language texts are essential components for analysis and treatment of Big Text data. Biomedical and health informatics are remarkable applications of natural language processing (NLP) and text classification in this field. In this study, we used machine learning techniques to perform the extraction of information related to surveillance of cancer. The surveillance of cancer prevalence and population-level statistics represent a key component to understanding the diseases and establishing treatment and prevention plans [4]. Cancer is one of the leading causes of death in America [2] [3]. However, despite more than one million new cases of cancer in the United States every year, human observers manually perform cancer surveillance. Such a manual process is both difficult to perform and possibly flawed. In order to address these manual classification challenges, it is desirable to develop mechanisms for automatic information removal from cancer text data.

Small feature sets that effectively characterise different disease states are an important use of genome-wide analysis of data on expression [5]. Patients with the same condition can have obviously different treatment responses and overall results in breast cancer. The strongest predictors like histological grade and the metastasized lymph node status fail to accurately identify breast tumours based on their clinical manifestations. It is reported that chemotherapy or hormone therapy could reduce the risk of distant metastases; however, in any event more than 70 percent of the patients receiving this medicine would have survived without this medication, and none of the methods reported currently allow patient-cut treatment policies [6].

Many methods have recently been suggested to categorise cancer sub-phenotypes into various risk groups to make sure that cancer patients receive appropriate therapy. Most classifiers reduce the area of the features by deriving compact features in a supervised or unattended way by selecting or extracting features [6]-[9]. Nevertheless, their performance is generally not scalable, and generally decreases sharply when used on data sets distinct from those used for construction of classifications. For example, two recent large-scale study gene expression profiles have selected a signature of 70 genes [6] respectively and another signature consisting of 76 genes [10] to predict distant metastases in breast cancer patients. These two studies achieved 0.7 accuracy in their own patient cohorts. However, when each method was applied to the data set of the other, it worked poorly with less than 0.55 accuracy [5].

We argue that the reasons why feature extraction based on the selected methods were so instable and independent from study are two fundamental reasons [5] [10]. A performance of features relies heavily on existing features and it is still difficult to detect the most appropriate features for the task.

The objective of this paper is to improve the performance in the prediction of cancer prognoses and develop a wider rating of results. To achieve this, a Bag of Words model bag for text feature extraction is deployed with certain improvements. The method proposed shows unsupervised learning of function in an optimal way through cancer document datasets compared to the previous feature selecting approaches.

The outline of the paper is presented below: section 2 provides the related works, section 3 discusses the proposed model for feature extraction. Section 4 evaluates the work and section 5 concludes the entire model.

2. RELATED WORKS

Li et al. [11] propose the latest Bag-of-Concepts (BoC) framework which automatically gains useful conceptual information from external knowledge and then probabilistically designs terms and phrases for a document into higher semantics (i.e. concepts). By utilising knowledge-based meaning content, BoC representation provides more semantic and conceptual text information, as well as improved human understanding interpretability.

Passalis and Tefas [12] developed a Bag-of-Embedded Words (BoEW) model which could represent text documents efficiently that would circumvent predominantly used methods such as the Bag-of-Words textual model. The proposal extends the traditional BoF model by introducing a weighting mask, which modifies the value of any learned codeword and optimises the model from bottom to bottom. The BoEW model also allows the learned presentation to be easily modified with unique input techniques to the information needs of the user.

Zagoris et al. [13] proposed the Model Bag Visual Words (BoVW) which seeks to distinguish and differentiate manuscripts from printed text machines. Initially, interest blocks are detected in the paper. A BoVW-based descriptor is determined for each block. The final characterisation of the blocks as handwritten, press-pressed or noise depends on the combination of binary SVM classification systems.

Sinoara et al. [14] proposed an approach to the representation of documents on the basis of embedded representations of terms and meanings. We combined the strength of the disambiguation of word meaning with the semantic richness of embedded vectors to construct embedded images of document collections. This approach leads to better and smaller representations.

Khan et al. [15] proposed a systematic system for the use of bagged discrete cosine transformers (BDCT) to provide offline identification for text-independent use. Universal codebooks are used for the first time for multiple predictor models. A final decision is then reached by using the majority voting rule of these predictor models.

3. METHODS

The diagnosis of breast cancer based on BoW is intended to extract the essential data and to address the undetected data obtained from the detection of breast cancer. The traditional BoW has three main steps:

Step 1: Extraction of feature.

Step 2: Getting the fundamental words. The clustering Algorithm namely Naive Bayes [17] is usually used to aggregate the features extracted into *k* clusters and the keywords are *k* clustering centers.

Step 3: Codebook building. The frequency of each fundamental word is counted as a word frequency vector in every cancer type document. A cancer-type description is given for each word frequency vector, and a codebook is provided for all word frequency vectors.

3.1 ADAPTIVE BASIC WORD SELECTION

Naive Bayes takes *k* as a parameter and aggregate all features in the step 2 of the construction of the traditional BoW model into *k* clusters *C*(*k*), where *k* must be manually adjusted. If *k* is too large, it means that there are too many basic words affecting classification efficiency. If *k* is too small, it is not enough to tell the number of basic words about different types of cancer. Thus, define SSE (summary of intra-class dispersion errors) as the following:

$$SSE = \sum_{i=1;x\in C_i}^{k} \left|x - m(i)\right|$$

where

x is considered as the sample present within a cluster *C*(*i*),

m(*i*) is considered as the cluster center of *C*(*i*).

SSE defines the intra-class dispersion stage, so smaller SSE means the clustering effect is stronger. SSE also decreases with the increase of *k*, but the decrease rate gradually decreases. The optimum number of fundamental words *k* is obtained when the descending curve of SSE tends to be flat.

3.2 STRUCTURE OF BOW MODEL

The BoW model is extended to three-layered structure in order to make it suitable for cancer diagnosis of planetary equipment under multiple operating conditions.

3.2.1 First layer:

- *Extracting features in real-time*: First, the characteristics are detected by the procedure and then the real-time values of each function point are calculated by the method.
- *Getting basic words in real-time*: Because the real time function is numeric, it is not necessary to cluster Naive Bayes. In addition, the distribution of real time values is relatively concentrated, so that every real time value is considered a fundamental word. The real-time calculation accuracy is to maintain two digits after the decimal point.
- *Building a real-time codebook*: The frequency of each basic word in each type of cancer is counted and the actual codebook is obtained.

3.2.2 Second Layer:

- *Extract other characteristics*: First, the functional points are detected by the method and then the other functional points values are calculated by the method.
- *Getting other fundamental words*: For the same reason, each other value is considered to be a basic word. The accuracy of calculating the other value is to retain two digits after the decimal point.
- *Build other codebooks*: The frequency of each other's basic word is counted for each type of cancer and the other codebook is obtained.

3.2.3 Third Layer:

- *Imbalanced features removal*: First, the feature points are detected using the method, and then the imbalanced feature of each point of feature is calculated using the method.
- *Getting balanced fundamental words*: The Naive Bayes algorithm combines all imbalanced features into k-clusters, and the basic words are k-clustering-centres. The optimal value of k is calculated according to the method.
- *Build a balanced codebook*: Each balanced basic word is counted in its frequency in each cancer type and a balanced codebook is obtained.

3.3 CANCER DIAGNOSIS METHOD

A document to be diagnosed is the method of cancer diagnosis.

- *Extraction of features*: First the feature points are found from the document to be diagnosed, and then real time, other and unbalanced characteristics are calculated for each feature point.
- *Approximation of basic words*: The features in the previous stage are replaced by the nearest word, other basic word and balanced fundamental word in real time, respectively.
- *Counting vector of word frequency*: The vectors of the document’s current, other and unbalanced word frequency are respectively counted.
- *First-layer cancer classification*: The distances between the document's real time word frequency vector and the vectors in the real-time codebook are calculated. The closer the distance is, the similar is the type of cancer represented by the vector in the actual codebook. The probability of the document type of cancer is calculated as follows:

$$P(i) = 1 - \left(\frac{d(i)}{\sum d(i)} \right)$$

where, N is considered as the cancer types, $P(i)$ is regarded as the probability or occurrence of a cancer type i and $d(i)$ is regarded as the distance between the cancer type i and word frequency vector in real-time codebook.

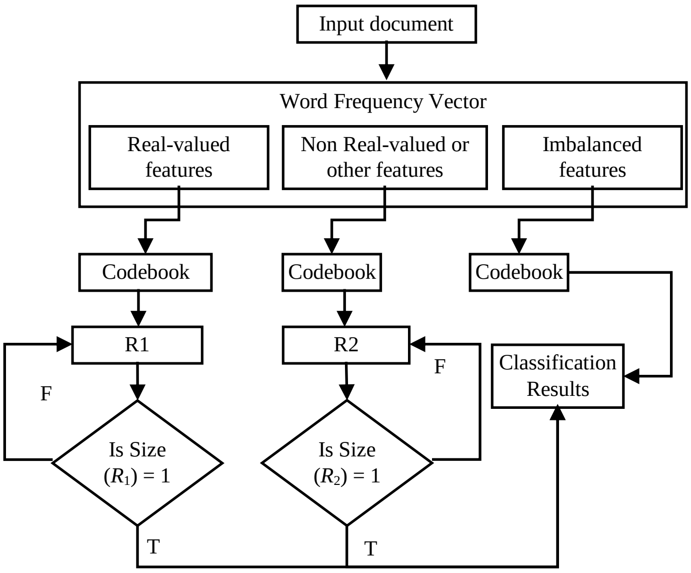

Fig.1. Proposed BoW concept

If $P(i)$ > 98%, R_1 of the first layer and R_1 of the second classification results are added to the classification result. If R_1 contains just one type of cancer, the final diagnosis will be made, otherwise it will be taken into the next layer.

- *Second layer cancer classification*: Calculate the distances from the document other word frequency vector to R_1 cancer types vectors in the other codebook, and calculate the probability of R_1 cancer types. For all i in R_1, when $P(i)$ > 98 percent, i is added to second layer R_2 classifications; and R_2 is also the prospective third layer classification results. If R_2 only contains one type of cancer, the final diagnosis results are taken, otherwise the next layer will be taken.
- *Third layer cancer classification*: The distances between the document Imbalanced word frequency vector and the vectors of the R_2 cancer types in the Imbalanced codebook are calculated and the type of cancer represented by the nearest vector is used for the final diagnosis.

After the word frequency vectors are obtained from the document to be diagnosed, the cancer diagnostic chart in Fig.1.

In the process of diagnosing cancer, the document being diagnosed first extracts real time, other and imbalanced characteristics. Then, each function is replaced by basic words such as the description in step 2 to make vectors for the word frequency. The distances between each type of word frequency vector and each type of code book are then calculated and the classification results R_1, R_2 and the final classification result are one or more types of cancer most likely. The probabilities of types of cancer are calculated are estimated using equation 2. If only one type of cancer exists in R_1 or R_2, the type of cancer is the final result of the classification.

4. PERFORMANCE EVALUATION

The proposed method is evaluated using a Wisconsin Diagnostic Breast Cancer [10] data collection from the University of California-Irvine Repository. The WDBC data sequence contained 569 samples including 357 benign and 212 malignant samples and a cell nucleus with 32 characteristics in 10 attributes. The data from the Wisconsin Breast Cancer Data Set are discussed in this report. Features are determined from a digital image of the breast weight fine needle aspirator (FNA).

Due to various dimensions and sizes of different features which can affect the results of the data analysis, normalisation should remove the effects of the dimensions and scales of the features. Features are also identical and errors can be avoided in large sizes. In particular, given that all samples in the WDBC data set contain labels not required to extract features in this analysis, the samples will be divided into two parts: 70% of training samples will be used in the extraction of features, and 30% of the samples are used for testing.

The feature include: area_mean, area_se, area_worst, compactness_mean, compactness_se, compactness_worst, concave points_mean, concave points_se, concave points_worst, concavity_mean, concavity_se, concavity_worst, diagnosis, fractal_dimension_mean, fractal_dimension_se, fractal_dimension_worst, id, perimeter_mean, perimeter_se, perimeter_worst, radius_mean, radius_se, radius_worst, smoothness_mean, smoothness_se, smoothness_worst, symmetry_mean, symmetry_se, symmetry_worst, texture_mean, texture_se and texture_worst.

Ten real-valued features are extracted from these features:

- Area
- Compactness ($p^2/(a\text{-}1)$), where p is the perimeter and a is the area.
- Concave points shows the total number of concave regions
- Concavity shows the severity of concave regions
- Fractal dimension uses coastline approximation

- Perimeter
- Radius is defined as the mean of distances from center to points on the perimeter
- Smoothness
- Symmetry and
- Texture

The study is evaluated in terms of various metrics over WDBC datasets and the results shows that the proposed BoW offered accuracy, f-measure, sensitivity, specificity, geometric mean and percentage error than existing text feature extraction methods that includes TF-IDF, word2Vec, BoW, Principle Component analysis (PCA) and Naive Bayes.

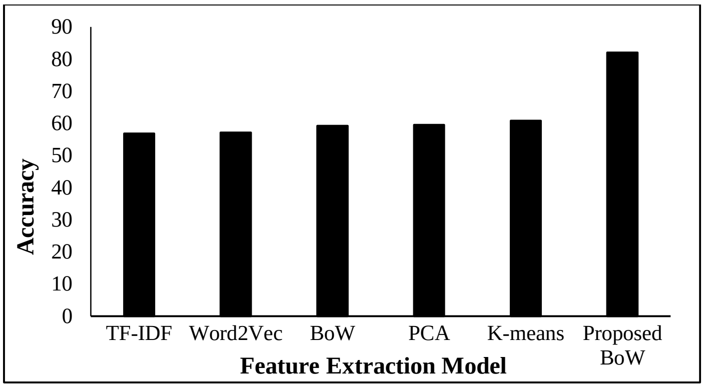

Fig.2. Accuracy

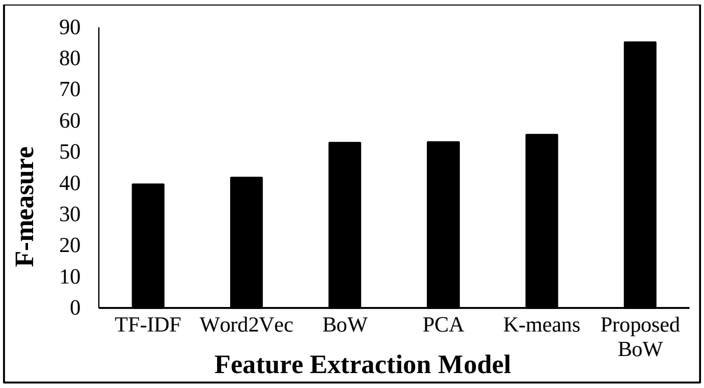

Fig.3. F-measure

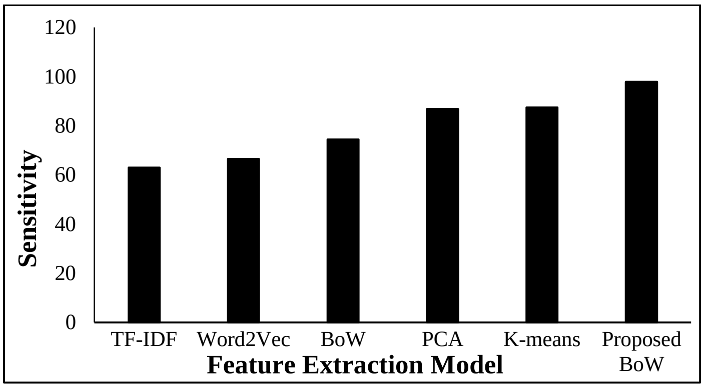

Fig.4. Sensitivity

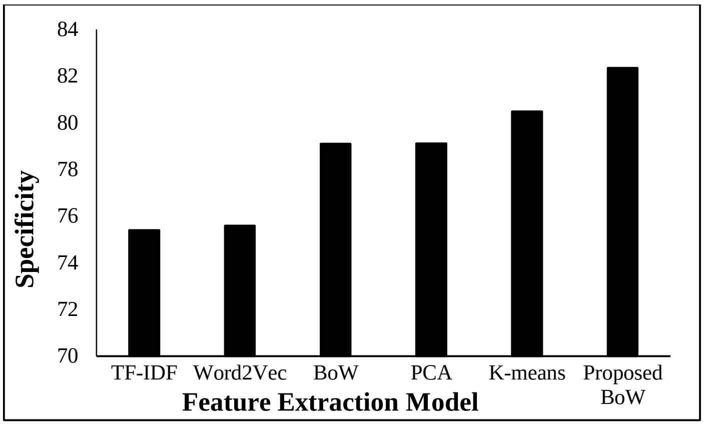

Fig.5. Specificity

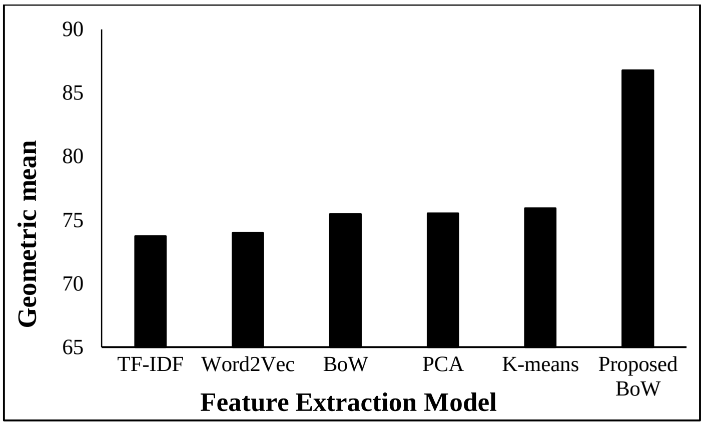

Fig.6. Geometric mean

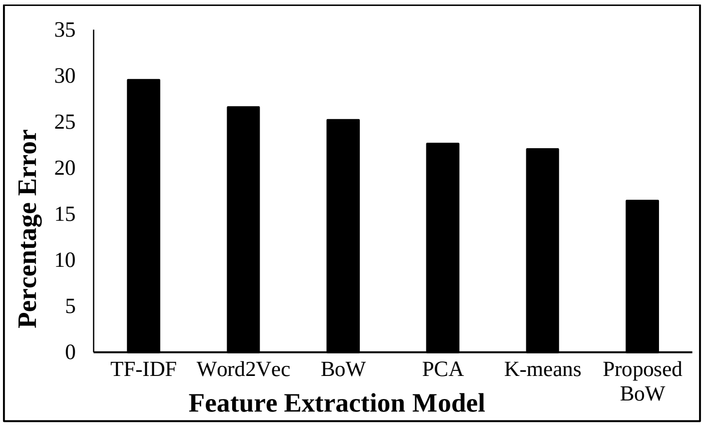

Fig.7. Percentage error

The study finds that the most appearing features includes area_se, area_worst, concave points_mean, concave points_worst, concavity_mean, concavity_worst, perimeter_worst, radius_mean, radius_worst, texture_mean and texture_worst.

From the results of Fig.2 – Fig.7, the results shows that the proposed BoW offered improved accuracy, f-measure, sensitivity, specificity, geometric mean and reduced percentage error than existing text feature extraction methods. The overall results shows that the proposed method obtains improved accuracy rate of 82% than existing methods with reduced percentage error of 16%.

5. CONCLUSION

In this paper, we present a new method with the enhanced Bag of Words extraction model to forecast the clinical outcomes of cancer patients. BoW presented more representative features from every text document for input cancer in three separate phases during the feature removal process. As shown by the test results, our methodology showed stronger prediction capabilities and better classifications based on in-depth learning techniques. Our analysis and discussion have shown that the properties automatically extracted by the BoW are able to easily generalise and increase the efficiency of functional extraction straight away. The future of cancer modelling should be explored in new methods to overcome the limitations described above. Better statistical analysis of heterogeneous datasets can yield more accurate findings and clarify disease outcomes. Further research is needed to create more public databases for the accurate collection of cancer data from all patients diagnosed. Their use by researchers would simplify their modelling studies and contribute to reliable findings and integrated clinical decision making.

REFERENCES

[1] D. Hanahan and R.A. Weinberg, “Hallmarks of Cancer: The Next Generation”, *Cell*, Vol. 144, No. 5, pp. 646-674, 2011.

[2] Priyadharshini Muthukrishnan, V. Sakthivel, Baskaran Ramachandran, K. Srihari and P. Balaji, “Querying Scalable and Redundant Data via. Recursive Queries in Orm Systems”, *Journal of Engineering and Applied Sciences*, Vol. 13, No. 7, pp. 5561-5564, 2018.

[3] American Cancer Society, “American Cancer Society Facts and Figures”, Available at: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2019/cancer-facts-and-figures-2019.pdf.

[4] M.J. Hayat, N. Howlader, M.E. Reichman and B.K. Edwards, “Cancer Statistics, Trends, and Multiple Primary Cancer Analyses from the Surveillance, Epidemiology, and End Results (SEER) Program”, *Oncologist*, Vol. 12, No. 1, pp. 1-14, 2007.

[5] W.K. Lim, E. Lyashenko and A. Califano, “Master Regulators used as Breast Cancer Metastasis Classifier”, *Proceedings of Pacific Symposium on Biocomputing*, pp. 504-515, 2009.

[6] Van De Vijver, M.J. He, L.J. Vant Veer, H. Dai, A.A. Hart, D.W. Voskuil and M. Parrish, “A Gene-Expression Signature as a Predictor of Survival in Breast Cancer”, *New England Journal of Medicine*, Vol. 347, No. 25, pp. 1999-2009, 2002.

[7] T. Sorlie, C.M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen and T. Thorsen, “Gene Expression Patterns of Breast Carcinomas Distinguish Tumor Subclasses with Clinical Implications”, *Proceedings of the National Academy of Sciences*, Vol. 98, No. 19, pp. 10869-10874, 2001.

[8] J. Li, A.E. Lenferink, Y. Deng, C. Collins, Q. Cui and E.O. Purisima, “Identification of High-Quality Cancer Prognostic Markers and Metastasis Network Modules”, *Nature Communications*, Vol. 1, No. 1, pp. 1-9, 2010.

[9] Y. Wang, J.G. Klijn, Y. Zhang, A.M. Sieuwerts, M.P. Look, F. Yang and T. Jatkoe, “Gene-Expression Profiles to Predict Distant Metastasis of Lymph-Node-Negative Primary Breast Cancer”, *Lancet*, Vol. 365, No. 9460, pp. 671-679, 2005.

[10] X. Zhou, J. Liu, X. Ye, W. Wang and J. Xiong, “Ensemble Classifier based on Context Specific Mirna Regulation Modules: A New Method for Cancer Outcome Prediction”, *BMC Bioinformatics*, Vol. 14, No. 12, pp. 1-6, 2013.

[11] P. Li, K. Mao, Y. Xu, Q. Li and J. Zhang, “Bag-of-Concepts Representation for Document Classification based on Automatic Knowledge Acquisition from Probabilistic Knowledge Base”, *Knowledge-Based Systems*, Vol. 193, pp. 105436-105444, 2020.

[12] N. Passalis and A. Tefas, “Learning Bag-of-Embedded-Words Representations for Textual Information Retrieval”, *Pattern Recognition*, Vol. 81, pp. 254-267, 2018.

[13] K. Zagoris, I. Pratikakis, A. Antonacopoulos, B. Gatos and N. Papamarkos, “Distinction Between Handwritten and Machine-Printed Text based on the Bag of Visual Words Model”, *Pattern Recognition*, Vol. 47, No. 3, pp. 1051-1062, 2014.

[14] R.A. Sinoara, J. Camacho Collados, R.G. Rossi, R. Navigli and S.O. Rezende, “Knowledge-Enhanced Document Embeddings for Text Classification”, *Knowledge-Based Systems*, Vol. 163, pp. 955-971, 2019.

[15] S. Kallam, R. Patan and A.H. Gandomi, “Improved Salient Object Detection using Hybrid Convolution Recurrent Neural Network”, *Expert Systems with Applications*, Vol. 166, pp. 1-23, 2020.

[16] WBDC Dataset Features, Available at: https://www.kaggle.com/quantumofronron/breast-cancer-data-set-feature-selection

[17] J. Zhang, C. Chen, Y. Xiang, W. Zhou and Y. Xiang, “Internet Traffic Classification by Aggregating Correlated Naive Bayes Predictions”, *IEEE Transactions on Information Forensics and Security*, Vol. 8, No. 1, pp. 5-15, 2012.