EasyChair Preprint

№ 7804

GL-LSTM Model for Multi Label Text Classification of Cardiovascular Disease Reports

Rim Chaib, Nabiha Azizi, Didier Schwab, Ibtissem Gasmi and Amira Chaib

April 18, 2022

GL-LSTM Model for Multi Label Text Classification of Cardiovascular Disease Reports

Rim Chaib[1,2] and Nabiha Azizi[1, 3] and Didier Schwab[4] and Gasmi Ibtissem[5] and Amira Chaib[6]

[1]Labged Laboratory of electronic document management
[2]Electronic department, Badji Mokhtar University, Annaba, Algeria
[3]Computer science department, Badji Mokhtar University, Annaba, Algeria
[4]LIG-GETALP Laboratory, Grenoble Alpes University, Grenoble, France
[5]Computer science department, Chadli Bendjdid University, El taref, Algeria,
[6]Biochemistry department, Badji Mokhtar University, Annaba, 23000. Algeria

chaib.rim23@gmail.com, azizi@labged.net

Abstract. In recent years, the rapid growth of electronic data and information has gotten a lot of attention, finding relevant information has become increasingly challenging. The goal of automatic text categorization is to classify textual articles based on their categories, especially in the medical domain. However, for some applications, objects must inherently be described by more than one label. In this research, a new scheme of medical multi-label text classification is investigated which is based on intelligent engineering features using GloVe technique and LSTM classifier. The main particularity of GloVe consists of extracting informative features to the word level automatically, allowing to capture global and local textual semantics. The choice of the LSTM model is motivated by the success that has been achieved in recent years; it allows to capture very long-term dependencies between words. The experiment of our approach named *GL-LSTM* on the cardiovascular text dataset has produced impressive results with an overall accuracy of 0.927 compared with related work existing in the literature.

Keywords: Multi-label classification, medical text, LSTM, GloVe, text categorization.

1 Introduction

With the vigorous development of technology and the proliferation of electronic documents from various sources access to relevant information at precious time has become difficult. Medical documents containing valuable information about patients. Medical texts are generally unstructured, complicated, and very difficult to manage [1]. Traditional methods of information mining cannot help the user to analyze and manage this big data. Hence, the organization and discovery of knowledge face new challenges as an important means of data mining using basic algorithms on machines to extract useful knowledge from these data [2]. So, it is convenient to use the automatic means to classify these documents to facilitate the task of finding information.

Automatic text classification is a main task of natural language processing which aims to classify electronic documents into one or more predefined categories [3]. The medical text can belong to several categories of text for example a medical report of a patient that can contain several diseases if we consider the diseases as categories, in this case, the classification task has become multi labels.

The extraction of relevant features increases the performance of the classification model therefore the extraction of informative features from medical texts can improve the development of the medical field and aid in clinical decision making.

Traditional methods of feature extraction depend mainly on prior knowledge. The use of these methods can generate redundant and insignificant features which could limit the performance of the classification model obtained [4].

Reading and understanding the information contained in millions of medical documents is a time-consuming process. One of the main problems in the classification of medical texts is how to extract the relevant features to better represent and classify these documents. So, it became necessary to adopt machine learning methods for automatic categorization of medical texts.

With the rise of deep learning and vector representation, natural language processing has seen an improvement in recent years. Recent studies in this area use Word embedding for the vector representation of words and the extraction of features at the word level [5-9]. Current state of the art shows the impact of feature extraction based on word embedding on text classification system performance [10,11]. Among the vector representation methods, we use Global Vectors for Word Representation. The GloVe technique is an unsupervised method of machine learning, it has established itself as one of the best natural language processing methods to generate digital vectors to represent text, and it has found success in numerous studies of text mining and natural language processing [6,10,12,13,20,21]. It allows us to capture local statistical information and global statistical information from a corpus.

Coming to the classification phase, many traditional machine learning algorithms have been used to solve the multi-label classification problem, but due to exponential growth of databases, these algorithms have some limitations.

The development of deep learning approach to help solve the big data problem in the multi-label classification case. Among the methods of machine learning are recurrent neural networks. *Long Short-Term Memory* is a recurrent neural network that relies on the memorization mechanism. This memory has the capacity to process long sequences and maintain a state for as long as necessary. LSTM also capable of learning long-term word dependencies and remembering data passed in memory.

In this paper, we propose an intelligent multi-label medical text classification system called *GL-LSTM* that relies on advanced machine learning techniques, using the GloVe model to extract relevant information at the word level. The LSTM classifier is adopted to classify these feature vectors. We also want to study the behavior of Long Short-Term Memory classifier on specific natural language processing tasks and to propose a learning method to improve performance.

The remainder of this paper is organized as follows: Section 2 presents related work. Section 3 describes the proposed methodology in detail. Section 4 deals with the discussions on the obtained experimental results. Finally, section 5 summarizes the proposed *GL-LSTM* and provides further research directions.

2 Related Work

Different approaches and techniques have been applied successfully to solve the multi-label classification problem. For example, in [14] used tree-based LSTM network proving that this model outperforms other recursive models in learning distributed representations of texts feelings. in [15] They proposed a multi-tag text classification model based on ELMo, they used a preformed word integration vector to extract features from text in order to solve the problem of the sentiment classification task. The objective work provided in [5] is the multi-label classification of electronic health records using a Feed-Forward neural network and several recurrent models based on the bi-directional GRU architecture with three variants of word incorporations. [16] Presents a methodology called HLSE for organizing labels and an analysis of the impact of using different combinations of Word Embedding model in the multi-label classification framework applied to the medical data set (MeSH), using a Deep Learning architecture. The authors of [17], proposed a neural network for multi-label document classification based on LSTM. They used two LSTMs, the first one consists in learning adaptive data representation by incorporating document labels, the latter are reorganized in function with a semantic tree, in which the semantics are appropriate for learning the LSTM. And the second performs the unified classification learning process by handling error propagation for a variable number of labels in each document. [18] created a multi-label classification system for different article documents and films using seq2seq / LSTM classifier. [19] called DNN classifier for multi-label classification of scientific papers using word embedding as a feature extraction method. [6] proposed a Classification of documents covering European Union laws, treaties, and other public documents, in this study they used the TF-IDF method to represent data and KNN classifier to perform the task of multi-label classification.

3 Methodology

The main objective of this work is to propose an intelligent system ensuring multi-label classification and categorization of textual biomedical data. In this section, we describe the general process of the proposed approach. This system goes through three stages after data preprocessing and preparation. The first step is the feature engineering by training GloVe model. The second step is the step of learning the classification model. And the last step is the test phase which evaluates the performance of the classification model. Fig.1 illustrates the architecture of the proposed *GL-LSTM* approach with the different steps. classification model.

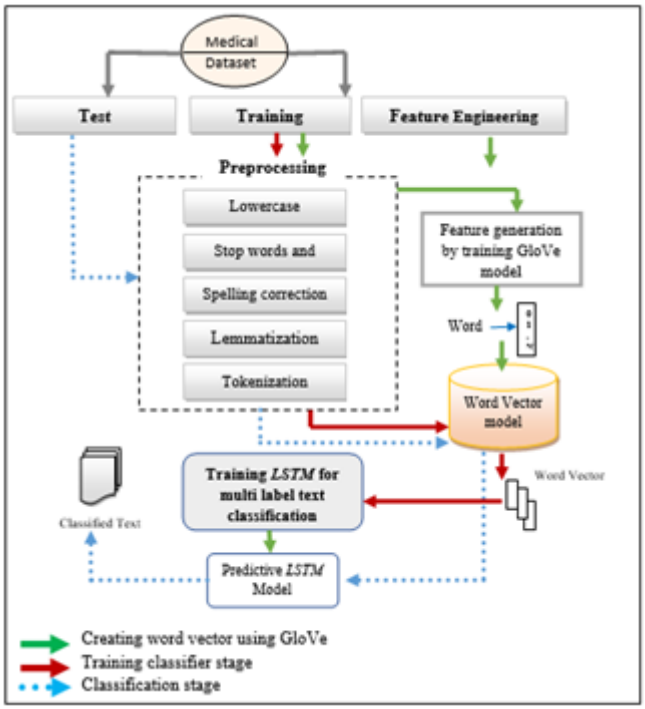

Fig. 1. Architecture of proposed approach.

3.1 preprocessing

The preprocessing step is a necessary task in text mining and natural language processing. To assess the effectiveness of the proposed approach for multi-label classification, we applied a set of basic preprocessing operations on the dataset, which are:

- Lowercase: It consists of transforming the input text data to lowercase. This step makes it possible to avoid having a multitude of copies for the same word. This preprocessing is applicable to most text mining and natural language preprocessing problems.
- Stop words and noise: Eliminate stop words which are common words in the language, punctuation, special characters, hashtags, HTML, URLs, redundant phrases and rarely used words have all been removed from the dataset. These words can interfere with the analysis of the text, for example the existence of hashtags in medical text can build noise because it characterizes a tweet.
- Spelling correction: This preprocessing step is to check the spelling of words and it will also reduce multiple copies of words. For example, "disease" and "desease" will be treated as different words even though it is the same word with a spelling error.

- Lemmatization: It consists of representing words in their canonical form.
- Tokenization: It aims to transform the text into a series of individual tokens. Each token represents a word such as "cardiovascular", "hypertension", "diagnosis", "physiology", and so on.

3.2 Feature Extraction

In this study an intelligent feature extraction technique based on deep representation, which is GloVe is investigated to analyze the behavior of the overall classification system. This method had considerable success in NLP. It makes possible to determine the relationship between words and to capture the global and local statistical information of a corpus. In this work a pre-trained GloVe model on Wikipedia is applied with vector size equal to 100. Therefore, the embedding matrix corresponding to our data will be created. The operation of the GloVe model consists in taking the log of the context matrix of word inclusions and then applying the matrix factorization to the context matrix to approximate the word embeddings. The architecture of the GloVe model is illustrated in Fig.2 which shows the factorization on the words of the context matrix to obtain the word vectors.

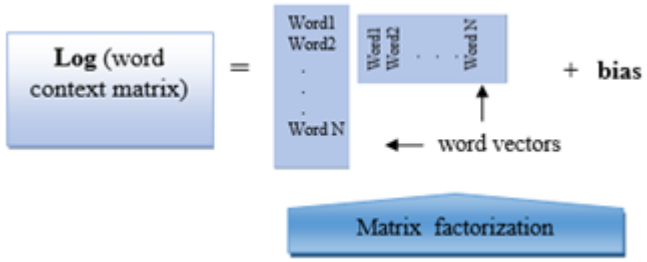

Fig. 2. Architecture of the GloVe model [11].

3.3 Classification Stage

The multi-label problem classification is a more challenging task comparing to classic problem classification where each instance is assigned to a single label. Multi-label classification consists of classifying each instance to one or more labels simultaneously so the number of label combinations for each instance increases, which makes the task of multi-label classification more complicated. In this study, the LSTM neural network is adopted to classify the features generated by the GloVe model. LSTM is the sequence processing model capable of learning long term word dependencies and remembering data passed in memory. The word vector produced by GloVe will be the input of the LSTM classifier. The LSTM layer is of size 128 which is connected to a dense layer of size equal to 23 neurons which is the number of classes in our dataset. The adopted LSTM architecture-based model is shown in Fig.3.

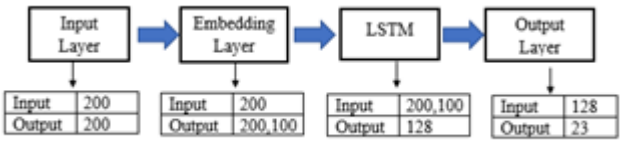

Fig. 3. Adopted LSTM architecture.

4 Experimentation

4.1 Data

The "Ohsumed" dataset is used to validate the *GL-LSTM* approach. This is a free online dataset from the MEDLINE database and consists of the medical summaries for 23 categories of cardiovascular disease. Ohsumed contains the first 20,000 documents among the 50,216 medical summaries for the year 1991. In this study a selected category subset available from [22] which has 13,929 documents. These documents are classified by categories. Fig.4 shows the distribution of documents by category.

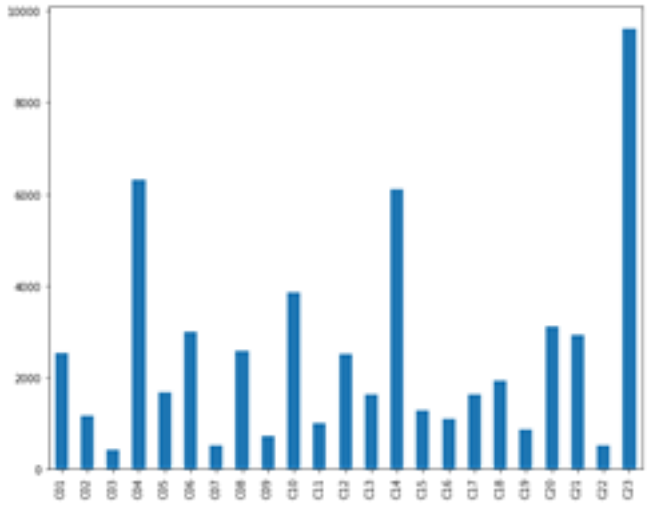

Fig. 4. The distribution of documents by category.

4.2 Evaluation and discussion

This section describes the different experiments that we carried out during this study. The initial goal of this work was to propose an intelligent system ensuring multi-label classification and categorization of textual Biomedical data. An advanced and intelligent feature extraction technique based on a deep representation is applied to extract the relevant information which is "Glove". As previously indicated that the GloVe method is a very powerful vector learning technique which has the advantage of al-

lowing us to qualify the relationship between words as well as to capture both the global and local semantics of a text to provide the vector of words.

Learning the GloVe model on our data set just gives one feature vector for each word. We used the pre-trained GloVe model with vector size set to 100. For example, the vector representation of the word 'diagnostic' produced by Glove is shown in Fig.5. After the preprocessing and vector representation steps, the data is divided into train data and test data to classify them. In the classification phase we want to study the behavior of Long Short-Term Memory classifier on specific tasks of automatic natural language processing and to propose a learning method to improve the performance of the proposed system while trying to better represent medical data.

```
embeddings_dictionary.get('diagnostic')

array([-0.54264  ,  0.41476  ,  1.0322   , -0.40244  ,  0.46691  ,
        0.21816  , -0.074864 ,  0.47332  ,  0.080996 , -0.22079  ,
       -0.12808  , -0.1144   ,  0.50891  ,  0.11568  ,  0.028211 ,
       -0.3628   ,  0.43823  ,  0.047511 ,  0.20282  ,  0.49857  ,
       -0.10068  ,  0.13269  ,  0.16972  ,  0.11653  ,  0.31355  ,
        0.25713  ,  0.092783 , -0.56826  , -0.52975  , -0.051456 ,
       -0.67326  ,  0.92533  ,  0.2693   ,  0.22734  ,  0.66365  ,
        0.26221  ,  0.19719  ,  0.2609   ,  0.18774  , -0.3454   ,
       -0.42635  ,  0.13975  ,  0.56338  , -0.56907  ,  0.12398  ,
       -0.12894  ,  0.72484  , -0.26105  , -0.26314  , -0.43605  ,
        0.078908 , -0.84146  ,  0.51595  ,  1.3997   , -0.7646   ,
       -3.1453   , -0.29202  , -0.31247  ,  1.5129   ,  0.52435  ,
        0.21456  ,  0.42452  , -0.088411 , -0.17805  ,  1.1876   ,
        0.10579  ,  0.76571  ,  0.21914  ,  0.35824  , -0.11636  ,
        0.093261 , -0.62483  , -0.21898  ,  0.21796  ,  0.74056  ,
       -0.43735  ,  0.14343  ,  0.14719  , -1.1605   , -0.050508 ,
        0.12677  , -0.014395 , -0.98676  , -0.091297 , -1.2054   ,
       -0.11974  ,  0.047847 , -0.54001  ,  0.52457  , -0.70963  ,
       -0.32528  , -0.1346   , -0.41314  ,  0.33435  , -0.0072412,
        0.32253  , -0.044219 , -1.2969   ,  0.76217  ,  0.46349  ],
      dtype=float32)
```

Fig. 5. Vector representation of the word 'diagnostic'.

LSTM is a recurrent neural network based on the memorization mechanism. This memory has the capacity to maintain a state for as long as necessary. The LSTM cell is also capable of modeling relationships and dependencies between words. After several empirical tests, we came to propose a recurrent neural network architecture which gave us good results. This neural network consists of an input layer containing 200 neurons i.e., the first 200 characters of each recording are taken, followed by an integration layer then an LSTM layer containing 128 neurons. The LSTM output layer is a fully connected layer made up of 23 neurons that calculate the abstract medical probabilities of input on 23 disease types (categories) with a sigmoid activation function for each neuron. Each neuron estimates the probability of belonging to a specific category with a value between [0.1]. After several empirical tests and changing of parameters, the best results obtained by *GL-LSTM* are presented in Table 1. The parameters adopted to obtain these results are batch size = 64, epochs=5, loss-function= binary_crossentropy and early stopping to regularize the machine learning model which avoids overfitting of model. The accuracy and loss diagrams for training model are also shown in Fig.6.

Table 1.

Model	Accuracy	Precision	F-Measure	Recall	Loss
GL-LSTM	0.928	0.906	0.891	0.887	0.230

From the experiments of this work, we can say that the proposed approach has improved the performance of multi-label classification system by comparing the results obtained with related work that uses the same 'Ohsumed' medical dataset.

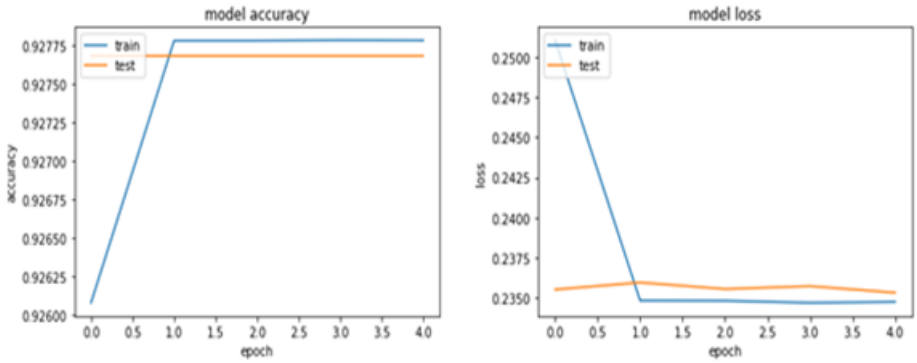

Fig. 6. The accuracy and loss diagrams.

Table 2 presents the results of related work with respect to the system proposed in this work.

Table 2.

Related work	**Used dataset**	**Feature extraction**	**Results**
[23]	Ohsumed	/	Acc=33.6
[24]	Ohsumed	Word embedding	Acc=70.45
[25]	Ohsumed	BOW	Acc=73.67
GL-LSTM	Ohsumed	GloVe	Acc=0.92

Based on our study, we can say that the performance of the classification system generally depends on how to represent the best feature vectors, which makes the learning task even more difficult.

5 Conclusion

Categorization of medical texts is a valuable area of text classification. In this paper, we have proposed an intelligent multi-label medical text classification system called

GL-LSTM that is based on advanced machine learning techniques. Extracting informative features from medical texts can enhance the development of the medical field and aid in clinical decision making. To accomplish this task, we used the GloVe method to extract the relevant information at the word level, it allows to calculate the global and local textual semantics of a text as well as the long-term syntactic regularities of a word. The Long Short-Term Memory classifier is adopted as the basic classifier for classifying medical texts. LSTM allows very long-term dependencies between words to be considered. The experimental results show the effectiveness of the proposed approach. Given the considerable success achieved by GloVe model and LSTM classifier, we want as perspectives to test other variants of LSTM for example: bidirectional LSTM as well as another feature extraction method like Elmo to better represent the medical data.

References

1. J. Lenivtceva, E. Slasten, M. Kashina & G. Kopanitsa, "Applicability of Machine Learning Methods to Multi-label Medical Text Classification", Computational Science – ICCS, 509–522, 2020.
2. R Wang, G. Chen, & X. Sui, "Multi label text classification method based on co-occurrence latent semantic vector space", Procedia computer science, 131, 756-764, 2018.
3. L. Lenc & P. Kral, "Word Embeddings for Multi-label Document Classification", Proceedings of Recent Advances in Natural Language Processing, Varna, Bulgaria, pp. 431–437, Sep 4–6 2017.
4. R.Chaib, N.Azizi, N.Zemmal, D. Schwab, & S. B. Belhaouari. "Improved Multi-label Medical Text Classification Using Features Cooperation", In *International Conference of Reliable Information and Communication Technology* (pp. 61-71). December 2020. Springer, Cham.
5. A. Blanco, O. Perez-de-Vinaspre, A. Pérez and A. Casillas, "Boosting ICD multi-label classification of health records with contextual embeddings and label-granularity", Computer Methods and Programs in Biomedicine 188, 2020.
6. W. Alkhatib, S. Schnitzer and C. Rensing, "Training-Less Multi-label Text Classification Using Knowledge Bases and Word Embeddings", International Conference on Knowledge Science, Engineering and Management, 2019. Springer, Cham.
7. R. Wang, W. Liu and C. McDonald, "Using word embeddings to enhance keyword identification for scientific publications", Australasian Database Conference, 2015. Springer, Cham.
8. J. Qiang, P. Chen, T. Wang and X. Wu, "Topic modeling over short texts by incorporating word embeddings", Pacific-Asia Conference on Knowledge Discovery and Data Mining,2017. Springer, Cham.
9. Y. Wang, S. Liu, N. Afzal, M. Rastegar-Mojarad, L. Wang, F. Shen, ... & H. Liu, "A comparison of word embeddings for the biomedical natural language processing", Journal of biomedical informatics, 87, 12-20, 2018.
10. A. L. Beam, B. Kompa, A. Schmaltz, I. Fried, G. Weber, N. Palmer, ... & I. S. Kohane, "Clinical concept embeddings learned from massive sources of multimodal medical data", In PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020 (pp. 295-306), 2019.

11. B. Chiu, G. Crichton, A. Korhonen, & S. Pyysalo, “How to train good word embeddings for biomedical NLP”, In Proceedings of the 15th workshop on biomedical natural language processing (pp. 166-174), August 2016.
12. P. Lauren, G. Qu, G.-B. Huang, P. Watta, & A. Lendasse, “A low-dimensional vector representation for words using an extreme learning machine”, 2017 International Joint Conference on Neural Networks (IJCNN), 2017.
13. M. Ibrahim, S. Gauch, O. Salman, M. Alqahtani, “An automated method to enrich consumer health vocabulariesusing GloVe word embeddings and an auxiliary lexical resource”, PeerJ Comput. arXiv preprint arXiv:2105.08812., 2021.
14. X. Zhu, P. Sobhani, & H. Guo, “Long short-term memory over tree structures”, arXiv:1503.04881, 2015.
15. W. Liu, B. Wen, S. Gao, J. Zheng and Y. Zheng, “A multi-label text classification model based on ELMo and attention”, MATEC Web of Conferences. Vol. 309. EDP Sciences, 2020.
16. F. Gargiulo, S. Silvestri, M. Ciampi and G. De Pietro, “Deep neural network for hierarchical extreme multi-label text classification”, Applied Soft Computing 79: 125-138, 2019.
17. Y. Yan, Y. Wang, W.C. Gao, B.W. Zhang, C. Yang and X.C. Yin, “LSTM 2: Multi-Label Ranking for Document Classification”, Neural Processing Letters 47.1: 117-138, 2018.
18. W.Chen, X. Liu, D.Guo, and M. Lu , “Multi-label Text Classification Based on Sequence Model”, Springer Nature Singapore Pte Ltd. 2019
19. F. Gargiulo, S. Silvestri and M. Ciampi, “Deep Convolution Neural Network for Extreme Multi-label Text Classification”, in Healthinf (pp. 641-650), 2018.
20. Y. Wang, S. Liu, N. Afzal, M. Rastegar-Mojarad, L. Wang, F. Shen, ... & H. Liu, “A comparison of word embeddings for the biomedical natural language processing”, Journal of biomedical informatics, 87, 12-20, 2018.
21. S. Dubois, N. Romano, D. C. Kale, N. Shah, & K. Jung, ”Effective representations of clinical notes”, arXiv preprint arXiv :1705.07025, 2017.
22. Available: http://disi.unitn.it/moschitti/corpora.htm Accessed 24 July 2020.
23. J. Camacho-Collados, & M. T. Pilehvar, “On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis”, arXiv preprint arXiv:1707.01780, 2017.
24. X. Liu, S. Wang, X. Zhang, X. You, J. Wu, & D. Dou, “Label-guided learning for text classification”, arXiv preprint arXiv:2002.10772, 2020.
25. B. Al-Salemi, M. Ayob & S. A. M. Noah, “Feature ranking for enhancing boosting-based multi-label text categorization”, Expert Systems with Applications, 113, 531-543, 2018.