Classwise Clustering for Classification of Imbalanced Text Data

K. Swarnalatha, D. S. Guru, Basavaraj S. Anami and Mahamad Suhil

Abstract In this paper, the problem of classification of imbalanced text data is addressed. Initially, imbalanceness present across the classes is reduced by converting each class into multiple smaller subclasses. Further, each document is represented in a lower-dimensional space of size equal to the number of subclasses using term-class relevance (TCR) measure-based transformation technique. Then, each subclass is represented in the form of an interval-valued feature vector to achieve the compactness and stored in a knowledgebase. A symbolic classifier has been effectively used for the classification of unlabeled text documents. Experiments are conducted on Reuters-21578 and TDT2 text datasets. The results reveal that the performance of the proposed method is better than the other existing methods.

Keywords Term-class relevance · Symbolic representation · Text classification · Hierarchical clustering

K. Swarnalatha (✉)
Department of Information Science & Engineering, Maharaja Institute of Technology Thandavapura, Mysuru, India
e-mail: swarnapradyu@gmail.com

D. S. Guru
Department of Studies in Computer Science, University of Mysore, Mysuru, India
e-mail: dsg@compsci.uni-mysore.ac.in

B. S. Anami
KLE Institute of Technology, Hubballi, India
e-mail: anami_basu@hotmail.com

M. Suhil
Department of Computer Science, GFGC, Paavagada, Tumukuru, India
e-mail: mahamad45@yahoo.co.in

V. Sridhar et al. (eds.), *Emerging Research in Electronics, Computer Science and Technology*, Lecture Notes in Electrical Engineering 545,
https://doi.org/10.1007/978-981-13-5802-9_8

1 Introduction

With the advancement of Web technology, the amount of text data over the Internet is increasing tremendously and the management of such a huge amount of text data is very difficult. Text classification, the task of automatically classifying unknown text documents into predefined categories, is an important task for most of the text management activities. Various applications such as news categorization, online marketing, e-mails, spam filtering, and text mining have also made the researchers to come out with efficient methods for text classification.

The three important factors which affect the efficiency of a text classification system are (i) the text can be either structured or unstructured, (ii) the number of features used to represent the text documents, and (iii) the number of documents present in different classes in the given collection. Numerous methods can be traced in the literature of text classification to handle these factors effectively [9]. Any machine learning approach for text classification requires a good representation to classify the text documents in an efficient way [23]. A most widely used representation scheme is the vector space model where a document is represented in the form of a vector of dimension equal to the number of terms present in the corpus. The conventional VSM is not effective as it leads to a very high-dimensional and sparse representation for the documents. Also, as the number of words increases the dimension of the representative feature vectors of the text documents also increases. But only a small subset of the entire population of features is helpful in achieving accurate classification. So, feature selection is a mandatory requirement to reduce the dimension through selecting only the important words for representing the documents.

In feature selection, the features are ranked based on some criteria which measure the importance of each feature so that a best subset of features can be chosen for the classification. The number of features to be used is usually fixed through empirical analysis. Some of the important feature selection methods available in the literature are terms-based discriminative information space [11], Fisher discriminant analysis [27], chi-square [26], information gain [1], term weightage using term frequency [18, 19, 22, 29], term frequency and document frequency-based feature selection [2], ontology-based [4], global feature selection methods [14, 15, 21]. In the literature of text classification, we can also trace a couple of methods which use transformation of features for dimensionality reduction [9], latent semantic indexing (LSI) [13], genetic algorithm (GA) [3].

Most of the methods available in the literature work well for balanced data collections and fail to perform well on imbalanced collections [15, 18, 29]. Few works focus on the classification of imbalanced text collections based on clustering methods, different type metrics for imbalanced text classification, clustering and dimensionality reduction [12].

2 Background and Motivation

Classification of imbalanced text collection is one of the challenging issues in designing an effective text classification system. To this end, [20] and [12] have recommended conversion of imbalanced text collection into a balanced one by dividing larger classes into a number of multiple smaller subclasses through classwise clustering. In their methods, once the subclasses are identified, it is recommended to treat each subclass itself as a class. Hence, the original K-class classification problem becomes a Q-class classification problem where K is the total number of classes originally present and Q ($\gg K$) is the total number of subclasses obtained due to clustering. Though both the methods are similar from the point of view of handling class imbalanceness, the approach in [20] has outperformed that of the [12]. In [12], the method uses a transformation approach which represents the documents in a K-dimensional space, where K is the number of classes present in the collection. While in [20], the dimensionality reduction is achieved through feature selection. Our observation here is that, once the natural groups among different classes are identified, representing the documents as feature vectors in K-dimensional space which is very small is not advisable as it is difficult to effectively capture the variation within each group effectively. Hence, a classifier may not generalize well on such a lower-dimensional space. With this motivation, in this paper once the original K-classes are converted into Q clusters, it is recommended to represent the documents in the form of Q-dimensional feature vectors by applying the same transformation applied in [12]. Since the value of Q is greater than that of K but not as large as the size of the bag of words, the classifier trained on a Q-dimensional vector space is expected to effectively capture the variations across different groups.

3 Representation of Documents in Lower-Dimensional Space

In the proposed method, it is recommended to represent the text documents in lower-dimensional space without using any explicit dimensionality reduction technique. The conventional bag-of-words-based representation leads to a high-dimensional vector space which is not suitable for effective classification without applying dimensionality reduction through an effective feature selection or feature transformation method. To overcome this difficulty, [10] proposed a text representation method which reduces the text document dimension from number of features n to a lower-dimensional space to the number of classes K. Further, the same method is adopted by Guru and Suhil [8] for better performance in text categorization through the introduction of a new term weighting scheme. The methods of [10] and Guru and Suhil [8] are used for the classification of imbalanced text documents. Hence, we have used the reduced representation of [10] along with the term-class relevance (TCR)

measure of Guru and Suhil [8] to represent the text documents in lower-dimensional form as explained below.

In this method, each document is initially represented in the form of a matrix F of size $\boldsymbol{n} \times \boldsymbol{K}$ as shown in Fig. 1, where $\boldsymbol{n}$ is the number of terms and K is the number of classes present in the corpus. The value of each location $F(i, j)$ is the weight of ith term t_i with respect to the class C_j computed using the TCR measure. Then, a feature vector f of dimension K is created as a representative vector of the document by aggregating the matrix $\boldsymbol{F}$ corresponding to the document such that $f(j)$ is the average relevancy of all terms present in the class C_j. Here, the dimension of the document is reduced to the number of classes K which is very small when compared to the original dimension n of the document. The TCR score for every term is calculated as

$$TCR(t_i, C_j) = Class\_weight(c_j) \times Class\_Term\_Weight(t_i, c_j) \times Class\_Term\_Density(t_i, c_j) \tag{1}$$

where

$$Class\_weight(C_j) = \frac{No.\ of\ Documents\ in\ Class\ C_j}{No.\ of\ Documents\ in\ Training\ Set} \tag{2}$$

$$Class\_Term\_Weight(t_i, C_j) = \frac{No.\ of\ Documents\ in\ Class\ C\ containing\ t_i}{No.\ of\ documents\ containing\ t_i\ in\ Training\ set} \tag{3}$$

$$Class\_Term\_Density(t_i, C_j) = \frac{No.\ of\ occurences\ of\ t_i\ in\ Class\ C}{No.\ of\ occrences\ of\ t_i\ in\ Training\ collection} \tag{4}$$

4 Proposed Method

The general architecture of the proposed model is given in Fig. 2. The different steps of proposed model are explained in subsequent subsections.

4.1 Clustering

It has been observed from the literature of text classification that most of the models work well for the balanced corpus. For an imbalanced corpus, the classes with large number of documents will generally dominate the classes with lower number of documents. One of the solutions for handling this issue is to convert the imbalance corpus into a balanced one. In Lavanya et al. [12] and Suhil et al. [20], it has been recommended to split a larger class into smaller subclasses by classwise clustering since the larger classes contain large intraclass variations. Hence, in this model we have converted large classes into small subclasses by applying hierarchical clustering

Terms in Document d	c_1	c_2	... ...	c_k
t_1	$w(t_1, c_1)$	$w(t_1, c_2)$	... ...	$w(t_1, c_k)$
t_2	$w(t_2, c_1)$	$w(t_2, c_2)$	... ...	$w(t_2, c_k)$
: :	: :	: :	: :	: :
t_n	$w(t_n, c_1)$	$w(t_n, c_2)$	... ...	$w(t_n, c_k)$
Feature Vector	$\frac{\sum w(t_i, c_1)}{n}$ $\forall i = 1....n$	$\frac{\sum w(t_i, c_2)}{n}$ $\forall i = 1...n$	... ...	$\frac{\sum w(t_i, c_k)}{n}$ $\forall i = 1...n$

Fig. 1 Representation of a document [10]

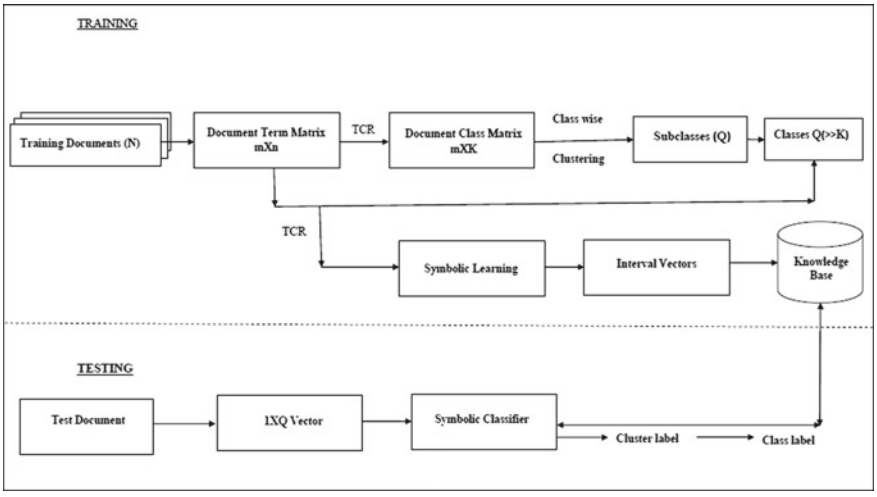

Fig. 2 Architecture of the proposed text classification method

technique. Finally, the clusters obtained due to all the K-classes are considered to be the classes and learning is applied to new classes.

Formally, let $\{C_1, C_2, C_3, \ldots, C_K\}$ be the K-classes present in the corpus. Each class C_j is converted into clusters of almost equal size by applying hierarchical clustering technique. Let $\left\{cl_1^j, cl_2^j, cl_3^j, \ldots, cl_{Q_j}^j\right\}$ be the Q_j number of clusters obtained for the class C_j. Similarly, let $\{Q_1, Q_2, Q_3, \ldots, Q_K\}$ be the number of clusters obtained for K different classes, respectively, and the total number of clusters is given by,

$$Q = \sum_{j=1}^{K} Q_j \tag{5}$$

The number of clusters varies from class to class which is based on size and intra-class variations present in the class. Since each cluster consists of similar documents, we can treat each cluster itself as a unique class. The representation scheme presented in Sect. 3 has been used to represent the documents in lower dimension since it is difficult to apply clustering in higher-dimensional space. By classwise clustering, we arrive at Q clusters which we treat as Q-classes, and hence, the original K-class classification problem is converted into a Q-class classification problem.

4.2 Representation of Documents of Each Cluster

Given a cluster Cl_j, we represent each document in the form of a Q-dimensional feature vector using TCR as explained in Sect. 3. The major difference here is that the TCR of each term is now recomputed with respect to each cluster. Each document in a cluster is represented in the form of a Q-dimensional feature vector $f = \{f_1, f_2, f_3, \ldots, f_Q\}$ which is very small when compared to the original dimension of the documents.

4.3 Creation of Knowledgebase of Interval-Valued Representatives

Recently, it has been shown that the approaches by the use of symbolic data outperform conventional algorithms in clustering and classification [7, 16]. Also, we can find in the literature some of the works on symbolic text representation and classification [5, 8, 12]. In our method, cluster-based interval-valued features are used for compact representation of documents to improve the performance on imbalanced text data.

Given a class C_j with Q_j number of clusters, each cluster cl_p^j consisting of m_p^j number of documents is represented by an interval-valued feature vector. We propose

to use interval-valued-type symbolic data to effectively capture the variations within a cluster of text documents. Another advantage of having such a representation is its simplicity in classifying an unknown document. Hence, the cluster cl_p^j is represented by an interval-valued symbolic representative vector R_{pj} as follows.

Let every document is represented by a feature vector of dimension K given by $\{f_1, f_2, \ldots, f_Q\}$. Then, with respect to every feature f_s, the documents of the cluster are aggregated in the form of an interval $[\mu^s - \sigma^s,\ \mu^s + \sigma^s]$ where μ^s and σ^s are, respectively, the mean and standard deviation of the values of f_s in the cluster. Hence, R_{pj} contains K intervals corresponding to the K features as,

$$R^{pj} = \left\{R_1^{pj}, R_2^{pj}, \ldots, R_Q^{pj}\right\}$$

where

$R_s^{ij} = [\mu^s - \sigma^s,\ \mu^s + \sigma^s]$ is the interval formed for the sth feature of the pth cluster cl_p^i of the jth class C_j. This process of creation of interval-valued symbolic representative is applied on all the Q clusters individually to obtain Q interval representative vector $\{R^{11},\ R^{12}, \ldots, R^{1Q_1}, R^{21},\ R^{22}, \ldots, R^{2Q_2}, \ldots, R^{K1},\ R^{K2}, \ldots, R^{KQ_K}\}$ which are then stored in the knowledgebase for the purpose of classification.

4.4 Classification

Given an unlabeled text document D_q, its class label is predicted by comparing it with all the representative vectors present in the knowledgebase. Initially, D_q is converted and represented as a feature vector $\left\{f_1^q, f_2^q, \ldots, f_Q^q\right\}$ of dimension Q as explained in Sect. 4.1. Then, the similarity between the crisp vector D_q and an interval-based representative vector R is computed using the similarity measure proposed by [6] as follows:

$$SIM\,(D_q, R) = \frac{1}{Q}\sum_{s=1}^{Q} SIM\,(D_q^s, R_s)$$

where

$$SIM\,(D_q^s, R_s) = \begin{cases} 1 & if\ (\mu^s - \sigma^s) \le f_s^q \le (\mu^s + \sigma^s) \\ \max\left[\frac{1}{1+abs((\mu^s-\sigma^s)-f_s^q)}, \frac{1}{1+abs((\mu^s+\sigma^s)-f_s^q)}\right] & otherwise \end{cases}$$

Similarly, the similarity of D_q with all the Q representative vectors present in the knowledgebase is computed. The class of a cluster cl which gets highest similarity with D_q is decided as the class of D_q as shown in Eq. (6).

$$ClassLabel(D_q) = Class(\arg\max_{i,j}(SIM\,(D_q, R^{ij}))) \tag{6}$$

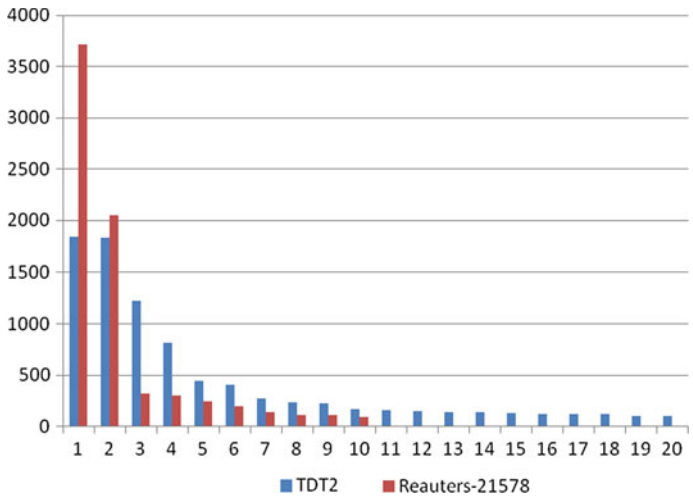

Fig. 3 Documents samples distribution in Reuters-21578 and TDT2 datasets

where R^{ij} is the representative of the jth cluster of the ith class.

5 Experimentation and Results

We have conducted the experiments to evaluate the method and to verify the efficiency of the proposed model by considering different training and testing sets. The performance of the proposed model has been evaluated using precision, recall, and F-measure in terms of both micro- and macro-averaging. The following sections present details about the imbalanced datasets considered for the experimentation and the results obtained.

5.1 Dataset and Experimental Setup

Dataset

To evaluate the efficiency of the model, we have conducted experiments on two benchmark imbalanced text datasets. The first benchmark dataset is Reuters-21578 which is collected from Reuters newswire, and we have considered a total 7285 documents from top 10 classes out of 135 classes with features 18,221 dimensions. The second dataset which is considered for our experiments is TDT2. A total of 8741 documents have been considered from the top 20 classes out of the 96 classes with 36,771 dimensions. Figure 3 shows the distribution of number of documents.

Experimental Setup
Experimentation has been conducted on each dataset by varying the percentage of training and testing from 10 to 80% in steps of 10% with 10 random trials each. The performance measures such as macro-precision, macro-recall, F-measure and the average performance of the 10 trials have been tabulated.

5.2 Results and Analysis

In this section, we present the results of the proposed method on both the datasets. The experiments were conducted by varying the number of clusters by varying the value of the inconsistency coefficient, and an optimal number of clusters is decided which produces the best results. More importantly to evaluate the goodness of the proposed model, a quantitative comparative analysis with the existing models is performed.

Table 1 and Table 2 show the results of the proposed method on Reuters-21578 and TDT2 datasets, respectively, in terms of macro-precision, macro-recall, macro-F-measure, and micro-F-measure. From these results, we can observe that the performance is increasing gradually with the increase in the percentage of training.

Table 1 Performance of the proposed model on Reuters-21578 dataset for 102 clusters

Percentage of training	Macro-P	Macro-R	Macro-F	Micro-F
10	0.6379	0.4016	0.4892	0.7959
20	0.7280	0.4899	0.5844	0.8229
30	0.7784	0.5726	0.6597	0.8406
40	0.7878	0.6272	0.6981	0.8553
50	0.7883	0.6689	0.7234	0.8624
60	0.7841	0.6967	0.7376	0.8645
70	0.7764	0.7205	0.7471	0.8693
80	0.7878	0.7443	0.7653	0.8743

Table 2 Performance of the proposed model on TDT2 dataset for 182 clusters

Percentage of training	Macro-P	Macro-R	Macro-F	Micro-F
10	0.6379	0.4016	0.4892	0.7959
20	0.7917	0.4343	0.5608	0.7602
30	0.8485	0.5405	0.6600	0.8003
40	0.8728	0.6095	0.7176	0.8284
50	0.8752	0.6517	0.7470	0.8437
60	0.8817	0.6809	0.7683	0.8505
70	0.8864	0.7193	0.7940	0.8628
80	0.8793	0.7337	0.7998	0.8621

To compare the performance of the proposed method with that of the available methods, we have selected two methods which try to handle the class imbalance by performing classwise clustering. The first method [20] uses classwise clustering for removing class imbalance and χ^2 for feature selection. The second method [12] uses classwise clustering for handling class imbalance and TCR for representation. In [12], each document is represented as feature vector of dimension equal to the number of classes originally present in the dataset, whereas in the proposed method, each document is represented by a feature vector of dimension equal to the number of clusters identified after classwise clustering.

Table 3 presents the comparison of the proposed method with that of Suhil et al. [20] and Lavanya et al. [12] in terms of macro-F and micro-F for both the datasets. The number of features used and the total number of clusters formed are also shown. It can be observed from Table 3 that the proposed method is better than the model of Lavanya et al. [12] in terms of both macro-F and micro-F. When it comes to the model of Suhil et al. [20], the proposed model has less performance. But the number of features used by Suhil et al. [20] is very high when compared to the number of features used by the proposed method. Thus, the model proposed by Suhil et al. [20] is very complex as it involves handling of very high-dimensional feature vectors.

Table 3 Comparison of the proposed method against the class-based method, Lavanya et al. [12] and Suhil et al. [20] with 70% training and 30% testing for Reuters and TDT2 datasets

Text corpus	Method	No. of features	Maximum no. of clusters formed	Macro-F	Micro-F
Reuters-21578	Classwise cluster+chi-square [20]	5000	453	0.8967	0.9351
	Cluster + TCR [12]	10	453	0.6843	0.8000
	Proposed method	102	102	0.7471	0.8693
TDT2	Classwise cluster+chi-square [20]	1000	339	0.9570	0.9631
	Cluster+TCR [12]	20	339	0.7600	0.8353
	Proposed method	182	182	0.7940	0.8628

6 Conclusion

In this paper, we have proposed a classwise cluster-based symbolic representation for imbalanced text classification using term-class relevance measure. To validate our results, we have conducted experiments with two different datasets, viz. Reuters-21578 and TDT2. The experimental results show that the proposed method works better with the class-based representation method. Hence, the classifier trained on a Q-dimensional vector space model can be used to capture the variations across different classes. In the future, the text classification can be conducted by this method using different dimensionality reduction techniques and clustering documents by considering various parameters like number of clusters and clustering technique.

References

1. Aghdam MH, Aghaee NG, Basiri ME (2009) Text feature selection using ant colony optimization. Expert Syst Appl 36(3):6843–6853
2. Azam N, Yao J (2012) Comparison of term frequency and document frequency based feature selection metrics in text categorization. Expert Syst Appl 39:4760–4768
3. Bharti KK, Singh PK (2015) Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering. Expert Syst Appl 42:3105–3114
4. Elhadad MK, Khaled M, Badran KM, Salama G (2017) A novel approach for ontology-based dimensionality reduction for web text document classification. In: International conference on information systems (ICIS)-2017, vol 978. IEEE, pp 5090–5507
5. Guru DS, Harish BS, Manjunath S (2010) Symbolic representation of text documents. In: Proceedings of the third annual ACM Bangalore conference (COMPUTE '10). ACM, New York, NY, USA, Article 18, 4 pp.
6. Guru DS, Nagendraswamy HS (2006) Symbolic representation of two-dimensional shapes. Pattern Recognit Lett 28:144–155
7. Guru DS, Prakash HN (2009) Online signature verification and recognition: an approach based on symbolic representation. IEEE TPAMI 31(6):1059–1073
8. Guru DS, Suhil M (2015) A novel term class relevance measure for text categorization. Procedia Comput Sci 45:13–22
9. Harish BS, Guru DS, Manjunath S (2010) Representation and classification of text documents: a brief review. IJCA Spec Issue on RTIPPR 110–119
10. Isa D, Lee LH, Kallimani VP, Rajkumar R (2008) Text document preprocessing with the Bayes formula for classification using the support vector machine. IEEE TKDE 20:1264–1272
11. Junejo KA, Karim A, Tahir MH, Jeon M (2016) Terms-based discriminative Information space for robust text classification. Inf Sci 372:518–538
12. Lavanya NR, Suhil M, Guru DS, Harsha SG (2016) Cluster based symbolic representation for skewed text categorization. In: International conference on recent trends in image processing and pattern recognition (RTIP2R)-2016, vol 709. Springer-CCIS, pp 202–216
13. Meng J, Lin H, Yu Y (2011) A two-stage feature selection method for text categorization. Comput Math Appl 62(7):2793–2800
14. Pinheiro RHW, Cavalcanti GDC, Ren TI (2015) Data-driven global-ranking local feature selection methods for text categorization. Expert Syst Appl 42:1941–1949
15. Pinheiro RHW, Cavalcanti GDC, Correa RF, Ren TI (2012) A global-ranking local feature selection method for text categorization. Expert Syst Appl 39:12851–12857
16. Punitha P, Guru DS (2008) Symbolic image indexing and retrieval by spatial similarity: an approach based on B-tree. Pattern Recognit 41(6):2068–2085

17. Rehman A, Javed K, Babri HA (2017) Feature selection based on a normalized difference measure for text classification. Inf Process Manag 53:473–489
18. Rehman A, Javed K, Babri HA, Saeed M (2015) Relative discrimination criterion—a novel feature ranking method for text data. Expert Syst Appl 42:3670–3681
19. Sabbaha T, Selamat A, Selamat MH, Fawaz S, Viedmae AEH, Krejcarg O (2017) Modified frequency-based term weighting schemes for text classification. Appl Soft Comput 58:193–206
20. Suhil M, Guru DS, Lavanya NR, Harsha SG (2016) Simple yet effective classification model for skewed text categorization. In: International conference on computing, communications and informatics (ICACCI)-2016. IEEE, pp 904–910
21. Uysal AK (2016) An improved global feature selection scheme for text classification. Expert Syst Appl 43:82–92
22. Uysal AK, Gunal S (2012) A novel probabilistic feature selection method for text classification. Knowl-Based Syst 36:226–235
23. Vieira AS, Borrajo L, Iglesias EL (2016) Improving the text classification using clustering and a novel HMM to reduce the dimensionality. Comput Methods Programs Biomed 136:119–130
24. Wang D, Zhang H, Li R, Lv W, Wang D (2014) t-Test feature selection approach based on term frequency for text categorization. Pattern Recognit Lett 45:1–10
25. Yang J, Liu Y, Zhu X, Liu Z, Zhang X (2012) A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. Inf Process Manag 48:741–754
26. Yang Y, Pedersen JO (1997) A comparative study on feature selection in text categorization. In: Proceedings of the 14th international conference on machine learning, vol 97, pp 412–420
27. Zeina D, Al-Anzi FS (2017) Employing fisher discriminant analysis for Arabic text classification. Comput Electr Eng 000:1–13
28. Zhang L, Jiang L, Li C, Kong G (2016) Two feature weighting approaches for naive Bayes text classifiers. Knowl-Based Syst 100(c):137–144
29. Zong W, Wu F, Chu LK, Sculli D (2015) A discriminative and semantic feature selection method for text categorization. Int J Prod Econ 165:215–222