Neural Computing and Applications (2022) 34:22519–22529
https://doi.org/10.1007/s00521-022-07669-5

ORIGINAL ARTICLE

ARTC: feature selection using association rules for text classification

Mozamel M. Saeed[1] · Zaher Al Aghbari[2]

Received: 19 January 2022 / Accepted: 22 July 2022 / Published online: 7 September 2022

Abstract
Feature vectors are extracted to represent objects in many classification tasks, such as text classification. Due to the high dimensionality of these raw feature vectors, the classification efficiency and accuracy are reduced. Therefore, reducing the size of feature vectors by selecting the relevant features that better represent the objects is an important aspect in text classification. Feature selection not only reduces the dimensionality of the feature vectors, but also produces more efficient classification models with higher predictive power. In this paper, we propose ARTC, which is an effective feature selection method that is based on the extraction of association rules to classify text documents. The extracted association rules discover the hidden relationships and correlations between the relevant words within the textual documents of a class and a cross different classes. Consequently, each class of documents is represented by a small set of contrasting features that are more effective in text classification. Our experiments show that ARTC outperforms other relevant techniques in terms of classification performance and efficiency.

Keywords Feature selection · Association rules · Text classification · Contrasting feature set · Text binary vector

1 Introduction

The increase in the number of online data necessitates the automatic classification of text. Classifying text is the process of labeling a given text based on the extracted features from the text [1]. The web is rapidly making most of the human knowledge available to us. Therefore, automatic text classification is gaining a lot of interest among researchers in the field of data mining. It is concerned with the effective and efficient use of natural language text processing and classification to maximize the usefulness of information extracted from the web. Text classification task is required in several applications such as spam detection [2], recommendation systems [3], sentiment analysis [4], and event detection [5].

Text classification methods map textual data into predefined classes that are useful to some application [6]. By mining these classes of document, interesting patterns that define each class can be discovered. However, due to the large size of these online documents, efficient feature selection methods that can be used in text classification are required.

Machine learning tools were utilized to classify documents based on their textual content. Such tools are commonly used by information retrieval systems (IRSs) to answer user queries. These machine learning tools require large number of labeled documents to learn the classification patterns. IRSs represent each textual document by a set of features. Typically, these feature vectors are highly dimensional that cause performance issues. The main issues are lower time performance and lower classification accuracy.

To improve the time performance and classification accuracy of the IRS systems, a feature selection method should be carefully designed to filter out the irrelevant features from the feature vectors. After the feature selection process, the feature vectors that represent the textual documents will be lower in dimensionality and contain the most effective features. Moreover, the selected features require less memory and will be processed more efficiently by a classifier.

Zaher Al Aghbari contributed equally to this work.

✉ Mozamel M. Saeed
m.musa@psau.edu.sa

Zaher Al Aghbari
zaher@sharjah.ac.ae

[1] Department of Computer Science, Prince Sattam Bin Abdulaziz University, Riyadh, Saudi Arabia

[2] Department of Computer Science, University of Sharjah, Sharjah, UAE

Data mining algorithms, such as association rules, can be utilized in feature selection. Association rules can discover relationships between apparently unrelated data items in large databases. Typically, association rules are used to find hidden associations or patterns between items in a large database. They are made up of an antecedent (head) and consequent (body), e.g., bread milk means that customers who buy bread (antecedent of the rule) are likely to buy milk (consequent of the rule) as well. Although this is an example of market basket analysis, the applications of association rules extend beyond that of the scope of market analysis.

Several methods have proposed text classification algorithms based on extracting association rules from the documents' feature space. These algorithms represent textual documents by feature vectors, where the features are the words in these documents. Furthermore, these methods extract association rules from the frequent features in those feature vectors. The main aim of these methods is to reduce the number of features that represent a text

document, which will lead to a reduction in the classification time. However, these feature vectors still produce relatively large number of features that represent a text document.

In this paper, we propose a robust feature selection method from text documents that filters out irrelevant features and retain the contrasting features. The proposed feature selection method leverages association rules to select the effective features for text classification. As shown in Fig. 1, the training set of text documents is preprocessed, where the text documents undergo noise cleaning, word stemming, and text structuring, and then each text document is represented by a binary vector. Next, the frequent words are extracted from the set of text documents of each class. That is, each class of documents is represented by a small set of contrasting features that are more effective in text classification. These frequent words are then used to generate representative association rules for each class of documents. The produced binary feature vectors representing the documents are smaller and yet more contrasting. In the phase of classifying unseen documents, an unseen document goes through the preprocessing steps similar to those of the training set. Then, the unseen document is represented by a binary vector, which is matched with the association rules of each class. Accordingly, the unseen text document is classified based on the majority vote.

This paper is divided into the following sections: In Sect. 2, we discuss the related work. In Sect. 3, we present the proposed ARTC method for finding association rules. We discuss the text classification in Sect. 4. In Sect. 5, we discuss the experiments and comparisons to peer methods. The conclusion is presented in Sect. 6.

2 Related work

Feature selection methods for text classification can be classified into three categories [8], wrapper, embedded and filtered. Wrapper methods employ a greedy search approach to compare all possible combinations of features in terms of classification accuracy [9]. Therefore, their computational complexity is high. On the other hand, the embedded methods are developed as part of the classifier [10], e.g., decision tree algorithm. A filtered method is typically a separate component of classification model. These methods are faster than wrapper methods since they do not have to undergo training. Therefore, our proposed feature selection method is catered to the filtered approach to be able to process large number of text documents efficiently.

The work in [11] presented a feature selection method that uses the relative document frequencies. It selects the features according to true positive rate and false positive rate. In [12], the authors presented a feature selection method that can extract effective memetic features. The work in [8] proposed feature selection techniques for employing new parameters to the relevance frequency methods. On the other hand, a feature selection method that utilized relevancy and redundancy of features is proposed by [13].

Text classification requires high computational power in the training phase of a classifier due to the high dimensionality of the textual documents. Therefore, reducing the dimensionality by feature selection methods of textual documents is of critical importance before training the classifier. Association rules method have been used to filter out the irrelevant features and extract meaningful ones. The method in [7] proposes a text classification algorithm, which uses association rules that are extracted using binary operations.

A vertical mining method to generate all candidate rules was proposed by [18]. The rules are ranked based on their confidence, support and number of features in the antecedent. Then, the highly ranked rules are selected. However, this method is computationally expensive due to the large number of generated candidate rules. To reduce the number of generated association rules, the method in [19] divided the training set into several categories equal to the number of expected classes. Then, they generated association rules from each category separately.

The approach in [14] is based on the heuristic that in a dataset containing textual documents that belong to one domain, relevant terms are likely to be associated with other relevant terms, while irrelevant terms are distributed randomly among documents. In [15], the paper identified several problems in IRS that may result in the wrong

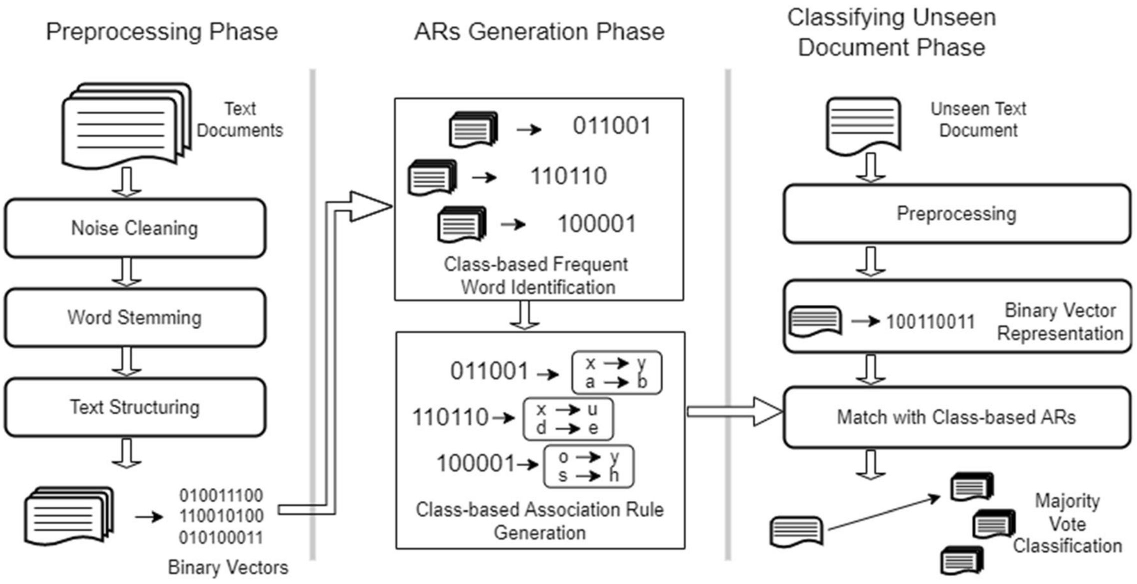

Fig. 1 ARTC workflow for classifying unseen documents

information being presented to the user. To address this, they offer SemanQE; a new semantic query expansion algorithm. It is made up of three components: a component for association rule-based query expansion, a feature selection component, and an ontology-based expansion component. The method in [16] reduced the dimensions of feature vectors by mining association rules. The association rules help to select the optimal set of features that influence the class category. Similar work was carried out by [17], whose experimental results also showed that association rules mining for feature selection dramatically reduces the cost for classification.

The method in [20] proposed a technique to reduce the number of generated association rules using Naive Bayes algorithm. To improve the classification of textual documents, [21] developed a compact classification model based on association rules. The model is based on evidentiary evidence and support, which represent the interestingness of the rules. On the other hand, [32] developed an educational data mining algorithm to build and visualize concept maps.

The idea of integrating fuzzy sets with association rule mining to classify text was proposed in [22] to improve the classification performance. The work in [23] extracted fuzzy association rules to represent the financial risk indicators. The method generated the fuzzy association rules using a parallel algorithm, where the rules meet the minimum credibility. Fuzzy association rules were also used to mine multi-search data from wireless sensor network data [24].

The method in [25] proposes an association rule-based classification model from imprecisely labelled data. The method represents training instances as belief functions. It argues that it can extract accurate and interpretable results. A similar method [26] is based on the integration of the following three modules: Apriori-based evidential association rule mining, rule pruning, and belief reasoning. The paper argues that the combination of these three phases create an accurate and compact classifier. In [27], a classifier based on association rule mining is proposed. The method reduces the number of extracted rules and integrates the confidence and support of the rules on the one hand and the class imbalance level on the other hand.

[28] proposes a novel methodology to mine association rules from distributed medical data sources (hospitals and clinics), where these medical data resources cannot be moved to other network sites. The desired global computation of association rules must be decomposed into local computations to match the distribution of data across the network. In [29], an explainable prediction model was developed using class association rules to identify veterans at risk of persistent post-traumatic stress disorder. The generated association rules serve as precursors to the long-term crisis in veterans. A model based on association rule mining [30] to explore the medical comorbidities of mental disorders is proposed. It was suggested based on the findings of the results that nurses should be provided with professional knowledge of comorbid conditions to better care for patients with mental illnesses. In [31], a model that consists of a combination of the association rules method and fuzzy soft set was proposed. The proposed model uses fuzzy soft set association rules to generate accurate classifiers for text documents

The methods presented above, which are summarized in Table 1, suffer from the high-dimensional feature vectors that represent text documents. This leads to a large number of association rules and therefore degrades the classification accuracy of unseen text documents. As shown in Table 1, most prior approaches suffer from high dimensionality of representative feature vectors. That is due to the large number of generated association rules. In this paper, we propose ARTC that represents a text document by a compact binary feature vector that led to a relatively small number of association rules. These rules are utilized to produce highly contrasting feature set for each class of documents.

3 ARTC method for finding association rules

Naturally, text data must undergo a sequence of preprocessing step before it can be ready for analysis tasks such as classification. For example, raw text has a large number of features, which is the number of words that represent the text. Such a large number of features degrade the performance of text classification. This is due to the fact that such a large number of features consist of irrelevant words to the topic. Moreover, the features of a text document may have high dependencies between them. Therefore, preprocessing of text documents is essential for effective classification. ARTC method preprocesses text documents of the training set in several steps as shown in Fig. 2: noise cleaning, word stemming and text structuring.

3.1 Noise cleaning

In the pre-processing phase punctuation, emojis, URLs, newlines, tabs, stop words, etc., are considered textual noise and irrelevant to the classification of textual documents. Also, HTML tags, if they exist, are removed since they do not contribute to the contrasting features. Therefore, ARTC's first preprocessing step is to remove such noise.

3.2 Word stemming

Words are used in their different forms in a text document depending on their grammatical context. These different forms of a word would greatly degrade classification of text. Therefore, a known technique in natural language processing (NLP) is called word stemming is used to solve this problem. Word stemming transforms the different forms of a word into their root, which is a unified representation of all forms. For example, a stemming algorithm might reduce the words *fishing*, *fished*, and *fisher* to the stem *fish*. Also, the stem need not be a word, for example the words *argue*, *argued*, *argues*, *arguing*, and *argus* are stemmed to *argu*.

Therefore, stemming is a type of normalization that standardizes some words into their stems. Although stemming results in loss of some of the original information of a word, it eases the processing a text document and improves its classification.

3.3 Text structuring

To structure text documents in the dataset, they are converted to binary forms, where each document is represented by a sparse binary vector. A binary vector that represents a certain text document is created by applying the following steps:

Table 1 Comparison between main related works

Paper	Data type	Document representation	Representative feature vector		
			Dimensionality	Method	Feature type
[11]	Text	Word vector	High	relative document frequencies	Frequent words
[12]	Text	Word vector	High	Feature filter + wrapper	Memetic
[8, 13]	Text	Word vector	High	Relevance frequency	Frequent words
[7]	Text	Binary vector	High	Association rules	Frequent words
[14–19, 27]	Text	Word vector	High	Association rules	Frequent words
[20]	Text	Word vector	High	Association rules + Naïve Bayes	Frequent words
[22–24, 31]	Text	Fuzzy word vector	Low	Association rules + fuzzy sets	Frequent words
[21, 25, 26]	Text	Fuzzy word vector	Low	Association rules + belief reasoning	Frequent words
[28–30]	Text	Word vector of clinical data	High	Association rules	Frequent words
ARTC	Text	Binary vector	Low	Association rules	Contrasting

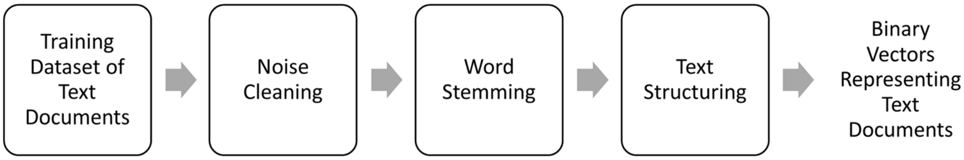
Training Dataset of Text Documents
Noise Cleaning
Word Stemming
Text Structuring
Binary Vectors Representing Text Documents

Fig. 2 Preprocessing phase of text documents

Table 2 Showing occurrences of words in text documents

DID	Sale	Unit	Unveil	Visit	Warner	Year	Bus. Class	Ent. Class	Pol. Class	Sprts Class	Tech Class
1	1	0	0	0	0	1	0	1	0	0	0
2	0	0	0	1	0	0	0	1	0	0	0
3	1	0	0	0	0	0	0	0	1	0	0
4	0	0	0	0	0	1	0	0	1	0	0
5	0	0	0	1	0	0	0	0	0	1	0

- Count the occurrences of every word in a document. As a result, a frequency list of all the words in the text document is generated.
- Remove the infrequent words. That is remove from the frequency list wordswhose frequency values are less than a specified threshold, called *support*, *s*, where *s* indicates whether a word is frequent in the text document or not. A frequency value of a word the is less than *s* indicates that the word is infrequent and thus cannot be used to represent the text document.
- A binary vector is created to represent all frequent words of a text document after noise cleaning and stemming. Words that are removed due to their low frequency value in the text document will be represented by a ”0” in the binary vector. On the other hand, words that are frequent will be represented by ”1” in the binary vector.
- The dataset of text documents is then represented in a table format (see Table 2), where each row is the binary vector that represents one text document. The column headers of this table are the different words in the dataset.
- A number of columns equivalent to the number of classes in the datasetare appended to the table. Each of these columns is labeled with the class name. For the purpose of creating a *training set*, each text document is labeled by a class it falls into by putting a ”1” in the column of the relevant class. However, the other class columns will be ”0” for this text document. Eventually, each class column C_col is a binary vector that indicates which text documents in the training set falls into this class.

4 Text classification

Classifying a text document into one of several predefined classes, requires the selection of a set of features that distinguishes the classes from each other. A more effective feature selection is finding a contrasting set of features for each class. The set of contrasting features for class C_i may not necessarily be the same set of features that best distinguishes class C_j. These distinguishing feature sets for the different classes, called contrasting sets, which have been applied by [9] on visual data, are discovered to represent the different textual classes.

In the proposed ARTC, the distinguishing feature sets are represented by the sets of frequents words for each class, as explained below. From these frequent words, ARTC generates a set of association rules for every class. The method generates the rules from the frequent words of a certain class, which means that the rules are the most distinguishing for this class. This process is depicted in Fig. 3.

4.1 Class-related frequent words identification

In this step, ARTC identifies the frequent words of the set of documents of a certain class. These frequent words effectively represent the class. That is, theseidentified frequent words have the most distinguishing power in predicting the class of a new text document.

To find the set of frequent words that represent a class, we perform a binary AND operation between each class column C_col and each of the frequent word columns. For example, the AND operation of the ”sale” word column

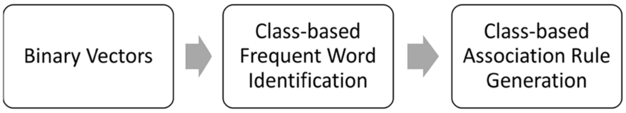
Binary Vectors
Class-based Frequent Word Identification
Class-based Association Rule Generation

Fig. 3 Association rules generation to represent each class

and ”Pol” class (see Table 2) would result in a ”1” for document ID number 3 (DID number 3) and ”0” under the rest of the DIDs. This makes the total of ”1”s of ANDing ”sale” with ”Pol” equal to one. After ANDing the class column C_col with the column of each word, the total of resulting ”1”s for each word is computed. The words with totals of ”1”s that is above a specified threshold are included in the contrasting set of the class.

On the other hand, those words whose total is below the threshold are discarded. That is these words are considered non-frequent in the documents of a class and thus not representative of the class. Thus, the process of ANDing a class column with all the words’ columns and then adding the results of ANDing for each word, the contrasting set of words (features) are determined for each class. Such contrasting set of words improve the ability of these words to distinguish the class they represent. Thus, each class will be represented by a set of contrasting words.

4.2 Creating rules

Once the contrasting set of a class has been identified, the association rules are generated from this set. To extract association rules from the contrasting set, subsets of word are generated. That is for each class, different subsets, or combinations, of the set of frequent words are generated. Then, the Apriori algorithm is applied to determine whether each of the generated subsets is frequent. That is each of the generated subsets is ANDed with the frequent words of each document (Table 2). If the ANDing operation results in a total of ”1”s that is greater than a certain threshold of generating a rule, then the subset is kept and will be considered for producing association rules.

Each kept frequent subset is used to produce an association rule, where the frequent word(s) in the antecedent of a rule and the class label in the consequent of the rule. Non-frequent subsets are discarded. The generated rules will then be used to predict that respective class. At the end of this step, a number of association rules will be generated for every class. Note that the number of generated association rules for each class will be relatively small. Consequently, the total number of association rules for all classes of text documents will be small. This leads to a low number of features representing each class.

4.3 Classifying unseen documents

As shown in Fig. 4, given an unseen text document, the same preprocessing steps are applied to the unseen document. Then, the unseen text document is represented by a structured binary vector by a similar process as that applied to the training set. The binary vector of the unseen document is fine-tuned by ANDing it with the association rules of each class. That is if a word in the rule is found in the unseen document, the corresponding cell of the binary vector of the unseen document remains ”1”, otherwise it becomes a ”0”. Then, the total of ”1” is computed, which represent the matching words between the binary vector of the unseen document and the rules of the class. This total value indicates the relevance $P\ (b_ik)$ of the unseen document to a class. The more matching rules of a class C_i to the unseen text document, the bigger the total value and thus the higher the relevance value $P\ (b_ik)$ between the unseen text document and the class C_i. Finally, the unseen document is classified to the class that gave the highest relevance value $P\ (b_ik)$.

5 Experimental results

ARTC was implemented using MATLAB. In this section, we discuss the dataset used and the experiments conducted to verify the feasibility of our approach.

5.1 Dataset

A dataset consisting of 2226 articles is collected from different sources such as Wikipedia, News agencies, sports websites, technology news, and business website. The collected articles belong to 5 classes: entertainment, politics, business, technology, and sports. The number of collected articles in each category are as follows: 328

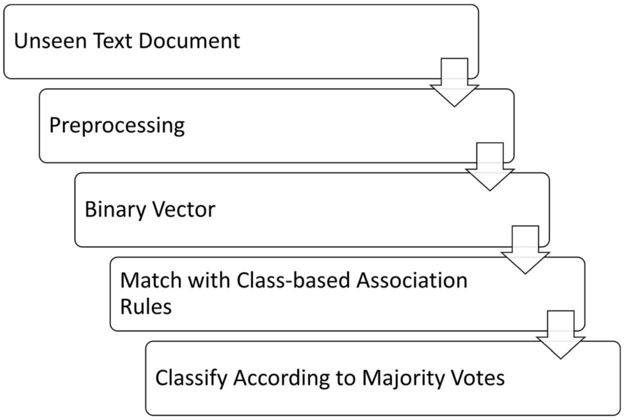
Unseen Text Document
Preprocessing
Binary Vector
Match with Class-based Association Rules
Classify According to Majority Votes

Fig. 4 Classifying unseen text documents

entertainment articles, 360 politics articles, 570 business articles, 370 technology articles, and 598 sports articles. These articles differ in size based on their source, where the size each document is in the range 60 to 600 words. The performance of ARTC was demonstrated by dividing the dataset into training and testing subsets. The training subset (70% of original dataset) is used to generate the association rules, while the testing subset (30% of the original dataset) is used to measure the accuracy of ARTC.

5.2 Effect of text preprocessing

The preprocessing step removes noise from raw text documents, such as stop words, punctuation, URL, emojis, etc. Then, it applies stemming to every word in the textual document. We show the effect of the preprocessing step through the below example. The example shows a raw document (text before preprocessing) and the same text after the preprocessing step. Note that noise has been removed and the words are stemmed. We used the Porter stemmer, which is in the NLTK package of python.

5.2.1 Sample text before pre-processing

Ad sales boost Time Warner profit Quarterly prof- its at US media giant TimeWarner jumped 76% to $1.13 bn (38223600 m) for the three months to December, from $639m year-earlier. The firm, which is now one of the biggest investors in Google, benefited from sales of high-speed internet connections and higher advert sales. TimeWarner said fourth quarter sales rose 2% to $11.1 bn from $10.9 bn. Its profits were buoyed by one-off gains which offset a profit dip at Warner Bros, and less users for AOL. Time Warner said on Friday that it now owns 8% of search-engine Google. But its own internet business, AOL, had has mixed fortunes.

5.2.2 Sample text after pre-processing

Ad sale boost Time Warner profit Quarterli profit at US media giant TimeWarn jump 76 to 113 bn 38223600 m for the three month to Decemb from 639m yearearli The firm which is now one of the biggest investor in Googl ben- efit from sale of highspe internet connect and higher advert sale TimeWarn said fourth quarter sale rose 2 to 111 bn from 109 bn It profit were buoy by oneoff gain which offset a profit dip at Warner Bro and less user for AOL Time Warner said on Friday that it now own 8 of searchengin Googl But it own internet busi AOL had ha mix fortune

5.3 Effect of representing text by binary vector

For each document, the word frequency list is computed. Words whose frequency value is below a threshold are discarded from the list. The threshold value indicates whether a word is frequent or non-frequent. Below, we show an example list of word frequency of a document.

{’giant’: 1, ’expectations’: 1, ’stake’: 3, ’rose’: 2, ’sales’: 4, ’fourth’: 3, ’performance’: 2,etc.}

Then, if the user sets the threshold to 2, for example, words whose frequency value below two are removed from the list as shown below.

{’stake’: 3, ’rose’: 2, ’sales’: 4, ’fourth’: 3, ’performance’: 2, etc.}

The frequency list of frequent words is converted into a binary vector. Each word in the frequency list, which is a frequent word, is represented as ”1” in the cell corresponding to that word. All other cells are filled with ”0”s. The binary vectors representing the documents in the training set are stored in a table (see Table 2), where each row represents one document.

In Table 2, we only show six words of the first five documents and hiding most columns for readability. Also, in Table 2, the five classes are indicated by the five right most columns, where a ”1” in column of a certain class C_i

indicates that the corresponding document belongs to C_i.

5.4 Class-related frequent words

To identify the frequent words that represent a certain class C_i, a binary AND operation between class column C_{col} of C_i and each of the frequent word’s column of Table 2. Table 3 shows the result of ANDing C_{col} of $C_{Business}$ and each of the frequent word’s column. Then, the sum of resulting ”1”s for each word is computed. The word with a sum of ”1”s that is above a specified threshold is added to the set of frequent words of the class. This set of words (features) is considered a contrasting set for each class. Table 4 shows an example of the frequent words that distinguishes each of the classes.

Table 3 Result of binary AND between frequent words and class column”Business”

Advert	Advertis	Advic	Advis	Aerospac	Affair	Affect
0	1	0	0	0	0	0
0	1	0	0	0	0	0
0	0	0	1	0	0	0
0	0	0	0	1	0	0
0	0	0	0	0	0	0

Table 4 Frequent words for every class

	Advert	Advertis	Advic	Advis	Aerospac	Affair	Affect
Business	0	1	0	1	1	0	1
Entertainment	0	0	0	0	0	0	0
Politics	0	0	1	1	0	1	0
Sports	0	0	0	0	0	0	0
Tech	1	1	0	0	0	0	1

5.5 Class-related association rules

To generate class-related rules, subsets of the frequent words are produced. Each of these subsets is then tested for becoming a rule by making the frequent words an antecedent of a rule and the class label becomes the consequent. In our dataset, there are 2840 possible frequent words. Among these, several hundred are marked as frequent for each class. Rules with different number of subsets are generated. In Table 5, we show the generate single-word subsets for every class that represent the possible subsets for the “Tech” class of the sample shown in Table 4. Due to the large number of possible subsets, we restricted our antecedent of the rules to only 2-itemset.

Next, we perform a binary AND between the generated subsets and the frequent words of each document. If the result is the same as the original subset, it is a match. However, if the result is greater than or equal to the threshold (which we have set to 2), then the subset is saved as a rule. The number of saved rules for each class is shown in Table 6.

Table 6 Generated association rules

Class	Number of rules
Business	44
Entertainment	32
Politics	51
Sports	78
Tech	63

5.6 Classifying unseen documents

We have split the dataset of text document into two subsets, 70% and 30% of the total size. Of the 2226 text documents, 1542 documents are used for training and the remaining 684 documents are used to test the performance of the proposed system. The test documents were preprocessed and converted into binary format.

An unseen document is binary ANDed with all the rules generated during the training phase. Then, the number of result matches between the binary vector of the unseen text document, b_i, and the rules of a certain class C_k indicate the probability, P (b_ik), of the unseen document being in the class. Therefore, we apply a voting algorithm to classify b_i based on the probability P (b_ik). The class with the highest value of P (b_ik) is assigned to the unseen document represented by b_i. That indicates the number of votes (number of rule matches) received by the unseen document b_i from the rules of class C_k. The following is an example of an unseen document text classified correctly as ’Business’:

Table 5 Possible subsets for”Tech” frequent words

Advert	Advertis	Advic	Advis	Aerospac	Affair	Affect
1	0	0	0	0	0	0
0	1	0	0	0	0	0
0	0	0	0	0	0	1

5.6.1 Sample unseen document

Peugeot deal boosts Mitsubishi Struggling Japanese car maker Mitsubishi Motors has struck a deal to supply French car maker Peugeot with 30,000 sports utility vehicles (SUV). The two firms signed a Memorandum of Understanding, and say they expect to seal a final agreement by Spring 2005. The alliance comes as a badly-needed boost for loss-making Mitsubishi, after several profit warnings and poor sales. The SUVs will be built in Japan using Peugeot’s diesel engines and sold mainly in the European market.

5.7 Classification accuracy

We tested ARTC using the collected dataset that belong to 5 different classes: business, entertainment, politics, sport and technology.

The average accuracy of the ARTC classifier was 82.2%. Meanwhile, the average precision was 86.9%, recall was 80.6%, and the F-Measure was 83.6%. Table 7 shows the confusion matrix of classifying the test documents to the five classes. As can be seen from Table 7, the test documents were classified with high true positives. Some false positives occurred due to the reason that some

Table 7 Confusion matrix

	Business	Entertainment	Politics	Sports	Tech
Business	**96.0%**	0.6%	2.5%	0.6%	1.3%
Entertainment	36.2%	**58.6%**	0.0%	1.8%	3.4%
Politics	18.7%	0.9%	**78.6%**	0.9%	0.9%
Sports	8.9%	1.2%	1.2%	**88.7%**	0.0%
Tech	13.3%	1.6%	2.3%	0.8%	**82.0%**

Bold values indicate the correct classification

text documents can be classified into more than one class. This is expected, since text documents usually discuss more than one topic.

Table 8 shows the precision P , recall R, F-measure $F1$, and accuracy ACC, which are measured by Eqs. 1, 2, 3 and 4, respectively, for each class.

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{1}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

$$F1 = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{TP} + \mathrm{FN}} \tag{3}$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{4}$$

where TP represents the true positives, FP represents false positives, TN represents true negatives, and FN represents false negatives.

ARTC was compared with the TF-IDF-based method for extracting association rules [33] in terms of performance. Particularly, we compared the average Precision, Recall, F1 measure, and Accuracy of both system on the whole dataset as shown in Figure 5. Note that ARTC outperformed the TF-IDF-based method due to the precise and compact representation of the binary vectors of their corresponding text documents.

We compared the proposed ARTC classification algorithm with the SVM-based classifier [34] as shown in Fig. 6. ARTC outperformed the SVM-based classifier since ARTC extracts contrasting set of features that have more distinguishing power in classifying the unseen documents.

Table 8 Precision, Recall, F-measure, and Accuracy for each class

	Precision	Recall	F-Measure	Accuracy
Business	61%	94%	76%	84%
Entertainment	92%	59%	73%	93%
Politics	91%	79%	83%	94%
Sports	96%	88%	91%	97%
Tech	93%	83%	87%	96%

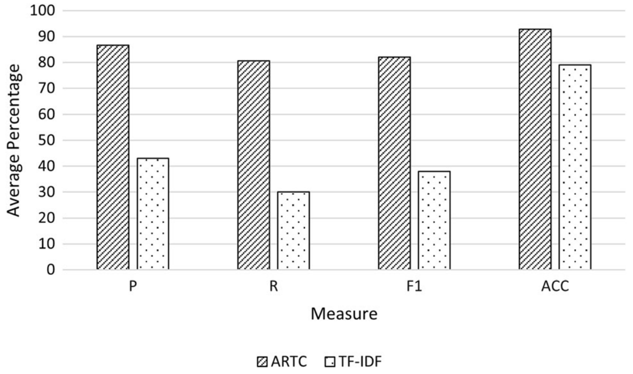

Fig. 5 Effect of feature representation on Precision, Recall, F-measure, and Accuracy for each class

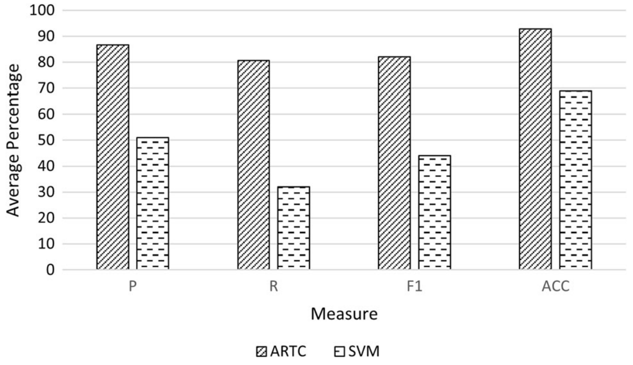

Fig. 6 Effect of classifier on Precision, Recall, F-measure, and Accuracy for each class

5.8 Discussion

The ARTC technique several advantages that can be summarized as follows:

- It represents each text document by a binary vector that corresponds to the frequent words of the text document.
- It leverages association rules to extracts a set of contrasting features (words) to represent the text document class of documents. This set of contrasting features is a small subset of the frequent words that represent the class of documents. That is achieved by removing the irrelevant features that do not contribute to the classification accuracy.

- Therefore, each class of documents is represented by binary feature vectors are small size and yet more contrasting, which leads to reducing classification time and improving the classification accuracy.
- It is independent of language. That is ARTC can be applied on feature vectors of frequent words representing text documents regardless of the language of the text.
- As compared to deep learning-based classification methods, ARTC is more memory efficient since it uses compact binary vectors to represent text documents. Moreover, deep learning-based methods are not interpretable since it is not possible to explain why one deep learning architecture is outperforming another.

On the other hand, ARTC has some limitations that can be summarized as follows:

- It does not take into consideration the semantic similarity between the frequent words. If taken into consideration, the binary feature vectors can be made more compact.
- It does not consider n-gram features since the features are single words. Considering n-gram features may improve the classification of text documents.

6 Conclusion

The proposed ARTC technique, initially, represents each text document by a compact binary vector that corresponds to the frequent words of the text document. A robust technique is developed to process the binary feature vectors of each class to extract a set of contrasting features, where this set is effectively small in size and yet improves the prediction of the class of an unseen text document. ARTC utilizes association rules mining to identify the effective distinguishing features of every class. The proposed ARTC method is independent of the language of the text documents. Moreover, ARTC learning time is quite fast since it only requires one pass through the training dataset to learn the rules. As compared to other peer systems, ARTC outperform them in terms of precision, recall, F1 measure, and accuracy. However, different language would require specific preprocessing to clean the documents and remove unnecessary content, such as stop words, digits, URLs, punctuations, etc. Once documents are cleaned by a language-specific pre-processor, ARTC can be easily applied.

Acknowledgements The authors have published the preliminary results of this work as a short version in a conference [35]. This version includes more detailed analysis of the proposed algorithms, and more experiments and comparisons.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Wang R, Chow C-Y, Kwong S (2016) Ambiguity-based multi-class active learning. IEEE Trans Fuzzy Syst 24(1):242–248
2. Makkar A, Garg S, Kumar N, Hossain MS, Ghoneim A, Alrashoud M (2020) An efficient spam detection technique for IoT devices using machine learning. IEEE Trans Industr Inf 17(2):903–912
3. Kanimozhi, U, Sannasi, G, Manjula, D, Arputharaj, K (2021) A user preference tree based personalized route recommendation system for constraint tourism and travel. Soft Computing, pp 1–20
4. Basiri ME, Nemati S, Abdar M, Cambria E, Acharya UR (2021) ABCDM: an attention-based bidirectional CNN-RNN deep model for sentiment analysis. Futur Gener Comput Syst 115:279–294
5. Peng H, Li J, Song Y, Yang R, Ranjan R, Yu PS, He L (2021) Streaming social event detection and evolution discovery in heterogeneous information networks. ACM Trans Knowl Discov Data (TKDD) 15(5):1–33
6. Cai J, Luo J, Wang S, Yang S (2018) Feature selection in machine learning: a new perspective. Neurocomputing 300:70–79
7. Sheydaei N, Saraee M, Shahgholian A (2015) A novel feature selection method for text classification using association rules and clustering. J Inf Sci 41(1):3–15
8. S¸ahin, D O, Kılı¸c, E, (2019) Two new feature selection metrics for text classification. Automatika 60(2):162–171
9. Al Aghbari Z, Junejo IN (2015) DisCoSet: discovery of contrast sets to reduce dimensionality and improve classification. Int J Comput Intel Sys 8(6):1178–1191
10. Uysal AK, Gunal S (2014) Text classification using genetic algorithm oriented latent semantic features. Expert Sys Appl 41(13):5938–5947
11. Kim K, Zang SY (2019) Trigonometric comparison measure: a feature selec tion method for text categorization. Data Knowl Eng 119:1–21
12. Lee J, Yu I, Park J et al (2019) Memetic feature selection for multilabel text categorization label frequency difference. Inf Sci 485:263–280
13. Labani M, Moradi P, Ahmadizar F et al (2018) A novel multivariate filter method for feature selection in text classification problems. Eng Appl Artif Intel 70:25–37
14. Webb GI (2007) Discovering significant patterns. J Mach Lear 68:1–33
15. Song M, Song IY, Hu X, Allen RB (2007) Integration of association rules and ontologies for semantic query expansion. Data Knowl Eng 63:63–75
16. Kaoungku, N, Suksut, K, Chanklan, R, Kerdprasop, K, Kerdprasop, N (2017) Data Classification Based on Feature Selection with Association Rule Mining. International MultiConference of Engineers and Computer Scientists, Hong Kong
17. Xie, J, Wu, J, Qian, Q (2009) Feature selection algorithm based on association rules mining method. Eighth IEEE/ACIS International Conference Computer and Information Science
18. Hadi WE, Aburub F, Alhawari S (2016) A new fast associative classification algorithm for detecting phishing websites. Appl Soft Comput 48:729–734
19. Alwidian, J, Hammo, B, Obeid, N (2020) Enhanced CBA algorithm based on apriori optimization and statistical ranking

measure. In Proceeding of 28th International Business Information Management Association (IBIMA) conference on Vision pp. 4291–4306

20. Hadi WE, Al-Radaideh QA, Alhawari S (2018) Integrating associative rule-based classification with naive bayes for text classification. Appl Soft Comput 69:344–356
21. Geng X, Liang Y, Jiao L (2021) EARC: Evidential association rule-based classification. Inf Sci 547:202–222
22. Fernandez-Basso C, Ruiz MD, Martin-Bautista MJ (2021) Spark solutions for discovering fuzzy association rules in big data. Int J Approximate Reason 137:94–112
23. Shang H, Lu D, Zhou Q (2021) Early warning of enterprise finance risk of big data mining in internet of things based on fuzzy association rules. Neural Comput Appl 33(9):3901–3909
24. Li, C, Li, W (2021) Automatic Classification Algorithm for Multisearch Data Association Rules in Wireless Networks. Wireless Communications and Mobile Computing, 2021
25. Geng X, Liang Y, Jiao L (2021) ARC-SL: association rule-based classification with soft labels. Knowl-Based Syst 225:107116
26. Geng, X, Liang, Y, & Jiao, L (2021) Evidential Association Classification for High-Dimensional Data. In 2021 IEEE 6th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), pp 100–105
27. Abu-Arqoub M, Hadi W, Ishtaiwi A (2021) ACRIPPER: a new associative classification based on RIPPER algorithm. J Inf Knowl Manag 20(01):2150013
28. Khedr AM, Al Aghbari Z, Al Ali A, Eljamil M (2021) An efficient association rule mining from distributed medical databases for predicting heart diseases. IEEE Access 9:15320–15333
29. Annapureddy, P, Franco, Z, Madiraju, P, Ahamed, S I, Flower, M, Hossain, M F, Winstead, O (2021) Identifying Precursors to Long-Term Crisis in Veterans Using Associative Classifier. In 2021 IEEE International Conference on Big Data (Big Data), pp 4633–4642
30. Wang CH, Lee TY, Hui KC, Chung MH (2019) Mental disorders and medical comorbidities: association rule mining approach. Perspect Psychiatr Care 55(3):517–526
31. Rohidin, D, Samsudin, N A, Deris, M M (2020) Association rules of fuzzy soft set based classification for text classification problem. Journal of King Saud University-Computer and Information Sciences
32. Shao Z, Li Y, Wang X, Zhao X, Guo Y (2020) Research on a new auto- matic generation algorithm of concept map based on text analysis and association rules mining. J Ambient Intell Humaniz Comput 11(2):539–551
33. Jabri, S, Dahbi, A, Gadi, T, Bassir, A (2018) Ranking of text documents using TF-IDF weighting and association rules mining. 4th international conference on optimization and applications, pp. 1–6
34. Puri, S, Singh, S P, (2019) An efficient hindi text classification model using svm. In Computing and Network Sustainability, Singapore, pp. 227–237
35. Al Aghbari, Z, Saeed, M, (2021) “Leveraging Association Rules in Feature Selection to Classify Text”, 4th International conference on Computer Networks and Inventive Communication Technologies, India.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.