Hindawi
Journal of Nanomaterials
Volume 2022, Article ID 9238968, 12 pages
https://doi.org/10.1155/2022/9238968

Research Article

Relevant-Based Feature Ranking (RBFR) Method for Text Classification Based on Machine Learning Algorithm

V. Durga Prasad Jasti,[1] **Guttikonda Kranthi Kumar**,[1] **M. Sandeep Kumar**,[2] **V. Maheshwari**,[2] **Prabhu Jayagopal**,[2] **Bhaskar Pant**,[3] **Alagar Karthick**,[4] **and M. Muhibbullah**[5]

[1]*Department of Computer Science and Engineering, VR Siddhartha Engineering College, Vijayawada, Andhra Pradesh 520007, India*
[2]*School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu 632014, India*
[3]*Department of Computer Science and Engineering, Graphic Era Deemed to Be University, Bell Road, Clement Town, 248002 Dehradun, Uttarakhand, India*
[4]*Renewable Energy Lab, Department of Electrical and Electronics Engineering, KPR Institute of Engineering and Technology, Coimbatore, 641407 Tamil Nadu, India*
[5]*Department of Electrical and Electronic Engineering, Bangladesh University, Dhaka 1207, Bangladesh*

Correspondence should be addressed to V. Durga Prasad Jasti; prasadjasti2018@gmail.com

Received 13 March 2022; Revised 12 July 2022; Accepted 21 July 2022; Published 3 August 2022

Academic Editor: Pandiyarasan Veluswamy

High dimensionality of the feature space is one of the problems in the field of text classification. Identification of optimal subset of features can optimize text classification process in terms of processing time and performance. In this paper, we propose a novel Relevant-Based Feature Ranking (RBFR) algorithm which identifies and selects smaller subsets of more relevant features in the feature space. We compared the performance of the RBFR against other existing feature selection methods such as balanced accuracy measure, information gain, Gini index, and odds ratio on 3 datasets, namely, 20 newsgroup, Reuters, and WAP datasets. We have used 5 machine learning models (SVM, NB, kNN, RF, and LR) to test and evaluate the proposed feature selection method. We found that the performance of the proposed feature selection method is 25.4305% times more effective than the existing feature selection methods in terms of accuracy.

1. Introduction

Massive amount of information is generated and pushed into the digital world every second through various sources such as web pages, blog contents, eBooks, social media contents, and review documents. As the content is increasing day by day, it becomes difficult to convert the content into an organized form which causes many problems such as difficult in searching and lack of summarization. Automatic text classification is one of the way to efficiently organize the documents. Supervised machine learning models such as support vector machines (SVM) [1], Naïve Bayes (NB) [2], k nearest neighbor (kNN) [3], random forest (RF) [4], and logistic regression (LR) [5] are very efficient in organizing content into one or more topics (or classes). There are wide applications of machine learning in the field of text classification such as spam detection [6], sentimental analysis [7], and topic classification [8].

There are three stages in text classification known as preprocessing, feature selection, and final classification. The preprocessing stage is responsible for formatting and removing useless words. Stop word removal, stemming, and text representations are few task performed in the preprocessing stage. Stop word removal eliminates useless symbols such as "is," "was," "that," and punctuation marks. Stemming is responsible for converting all the derived words into its root form (e.g., "running" is converted to "run," and "walked" is converted to "walk"). Word representing formats the document into usable text. Features are identified in this stage. There are many text representations such as Bag-of-Words (BoW) [9] and n-gram [10].

A feature is the indivisible atomic unit in a text document. A text corpus may contain many documents $D=\{d_1, d_2, d_n\}$. Each document contains m number of unique features, and the entire text corpus contains k number of unique features such as $F=\{f_1, f_2, \cdots, f_k\}$. As the number of documents increases, the corresponding feature size also increases which increases the classification complexity, increases time, and decreases the accuracy. Hence, an optimal subset of F should be found to represent the document much better and increase the classification performance. The total number of subset possibility is 2^k-1 (excluding the null set), so it is not practically possible to brute force all the combinations; thus, there are various feature selection algorithms which are aimed at finding out the optimal combinations in much easier way.

There are three types of feature selection methods known as filter based, wrapper based, and embedded based [11]. Filter-based methods are model independent which picks the features based on statistical methods like correlation and chi-square. Filter-based methods are faster than the other two types but it cannot identify the dependency between the features. Wrapper-based methods are model dependent that means for each model, separate sets of features are selected. Wrapper-based methods use an evaluation strategy to pick the optimal subset. The embedded-based method combines both the filter based and wrapper based. Wrapper-based methods inherit both the positives and negatives of filter and wrapper based.

In this paper, we propose a filter-based feature selection method called as Relevant-Based Feature Ranking (RBFR) algorithm which identifies the most important features and removes irrelevant features from the feature space. The proposed method first ranks all the features according to two metrics known as true positive rate (TPR) and false positive rate (FPR). Then, the features from top TPR are picked; within the chosen list, the features with high FPR are removed. The list is appended by the common features selected by odds ratio (OR), information gain (IG), and chi-square feature selection methods. We have compared the proposed method with well-known standard feature selection methods such as balanced accuracy, OR, IG, and Pearson correlation. The main contributions are listed as follows:

(i) To develop a filter-based feature selection method which is able to pick the most important features that could describe the target class better

(ii) To identify and eliminate overlapping or weak features that poorly represent the target class

(iii) To utilize the merits of other filter-based methods to pick correct features

The above-mentioned contributions are aimed at picking the high rich features that could represent the target class better than the other features; additionally, the error in the selected features should be identified and removed to increase the performance. Moreover, the high features selected by other filter-based methods are also utilized in the feature selection process.

The rest of the paper is organized as follows. Section 2 briefs the literature related to feature selection. Section 3 contains the working of the proposed algorithms. Section 4 presents the experimental results and the comparison with existing machine learning models and with other existing works. Finally, the conclusion is present in Section 5.

2. Related Works

In this section, we brief the recent works in the field of feature selection in text classification and list out the comparison, merits, and limitations.

A research work done by [12] proposes a feature selection method that uses correlation between each feature to the class. They have strengthened the positive features and weakened the negative features. A margin-based feature selection is implemented to increase the performance of the classification. They have evaluated their proposed filter-based method in thirteen datasets and showed the superiority over existing feature selection methods.

Feature selection can also be done in many stages. A work by [13] proposes a three-stage feature selection. In the first stage, they have incorporated particle swarm optimization to search for optimal features in the feature space. The second stage, the redundant features are found and removed from the selected features. The last stage is used to measure each feature for their significance; if the measure is too low, they are deleted from the feature space. Thus, one stage for selecting the features and two stages for removing irrelevant features are used.

The feature selection proposed by [14] focuses on selecting features in two decision levels. In the first level, they have used learners to find the relevant features. The filtration of learners is done to find the high confident learners. The elected learners are allowed to vote in the second level to pick the most relevant features among the feature space.

Clustering is used for grouping features and picking the relevant features in a work proposed by [15]. The redundancy and relevancy problems are solved by the clustering algorithm. A sorting algorithm is used which arranges all the features in the clustering space. Correlation is the main metric used in the sorting algorithm to rank all the features.

An embedded based feature selection was proposed by [16] for classification on Twitter review. As it combines both filter and wrapper methods, it eliminates the semantic problem. Transfer learning is used along with filter-based methods such as information gain, Pearson's correlation, and wrapper-based methods such as expectation maximization. A weight-based deep learning model is implemented to test the performance of the proposed method.

The irrelevant and redundant features present in the text corpus create a negative impact in text classification. A hybrid filter-based feature selection introduced by [17] combines principal component analysis and information gain. In their experiment, they found that their proposed feature selection method reduces the dimension of data significantly by picking the correct feature subset thus reducing the training time.

A comparison of feature selection was done by [18]; they have used seven filter-based methods, two wrapped-based methods, and one embedded-based method to test the significance of the classification. Three models artificial neural network, support vector machine, and random forest were used in their experiment. Several combinations of feature selection and classifiers are made, and the most appropriate subset is found based on the training performance.

Instance selection is the method of selecting/removing instance. Reducing the number of instances is also one of the methods to increase the performance of the classification. Ensemble methods are also popular in feature selection such as in [19] where the authors have used both feature selection and instance selection. Three-feature selection algorithms along with instance selection are used in their experiment. Two ensemble-based techniques are used in the experiment.

Redundancy and dependency identification is generally good in filter-based methods [20]; a work [21] shows that mutual information feature selection is effective in finding correlation between the features and the target class. When it comes to the fuzzy-based environment, the mutual information like other filter-based methods is weak in calculating correlation and dependencies. They adopted a fuzzy independent classification on a fuzzy-based data space; then, based on the proportion of classification error, they adjust the fuzzy-based feature selection.

Feature selection is optimized by using genetic programming as mentioned in [22]. A hybrid feature selection is done by merging multiple filter-based feature selection methods. A feature construction algorithm is utilized to optimize the selected features. Nine datasets were used in their experiment, and the comparison shows that the feature construction algorithm is effective (Table 1).

From the above-mentioned literature, the feature selection needs lots of improvement, especially when considering the relevancy. Thus, we propose a feature selection which is able to extract the relevant features which improves the efficiency of the text classification.

2.1. Few Existing Feature Selection Methods. This section presents an overview of three popular feature selection filter-based methods.

2.1.1. Information Gain. Information gain [28] is a supervised feature selection methods which is used to rank the feature according to the word's contribution based on its presence or absence in a particular set of text inputs [29]. IG is calculated as

$$\begin{aligned} \mathrm{IG} = &-\sum_{i=1}^{m} P(c_i) \log P(c_i) + P(t) \sum_{i=1}^{m} P(c_i|t) \log P(c_i|t)\, P(\bar{t}) \\ &\cdot \textstyle\sum_{i=1}^{m} P(c_i \mid \bar{t}) \log P(c_i \mid \bar{t}), \end{aligned} \tag{1}$$

where m represents the total number of target classes. If binary classification is used then m value is 2. $P(C_i)$ denotes the probability of class *i*. $P(t)$ is the probability of the word t when t is present in the document, and similarly, $P(\bar{t})$ represents the probability of the word t when t is absent in the document. $P(c_i \mid t)$ and $P(c_i \mid \bar{t})$ are the conditional probabilities.

2.1.2. Chi-Square. Chi-square [30] is the test of independence of a feature with the target class. It is used to measure how much a term is diverged from its dependent class [31]. CHI is calculated using the formula shown as follows:

$$\begin{aligned} \mathrm{CHI} = {} & t\big(tp, (tp + fp) + \mathrm{ve}_{\mathrm{prob}} + t\big(fn, (fn + tn) + \mathrm{ve}_{\mathrm{prob}}\big) \\ & + t\Big(fp, (tp + fp) - \mathrm{ve}_{\mathrm{prob})} + t(tn, fn + tn) - \mathrm{ve}_{\mathrm{prob}}\Big). \end{aligned} \tag{2}$$

The symbols $+\mathrm{ve}_{\mathrm{prob}}$ and $-\mathrm{ve}_{\mathrm{prob}}$ represent the probability of the positive class and the negative class, respectively.

2.1.3. Pearson Correlation. Pearson correlation is one of the good statistical measures to test the dependence of a feature towards the target class [32]. It is unaffected by overfitting [33]. It is calculated by the formula as described as follows:

$$\mathrm{PC} = \frac{\mathrm{Cov}(X, Y)}{\sigma X \sigma Y}. \tag{3}$$

The existing feature selections have lots of problems such as lack of representation of class unique features, problems in removing the unless and common features, and unable to perform negativity test.

2.2. Overall Drawbacks in Existing Feature Selection Methods. Feature selection is done to reduce the dimensionality of features in the dataset. Good features need to be identified to separate the classes. As the number of features increases, the complexity of the classifier is also increased; this creates a need for better feature selection methods [34].

Most existing feature selection methods use a weighted method such as frequency and distribution; these feature selection methods fail to pick the class unique features; that is, when one feature is very specific to one class or few classes, that feature is very important for a classifier to determine the class as the classifier feels very easy to identify the class.

Another problem in the feature selection is many methods rely on positive test; that is, if a feature is present, then an appropriate class can be identified; however, negativity test is also one of the powerful methods to eliminate weak candidates in the classification. There are only limited methods for the negativity test.

Combining two or more feature selection methods lets the classifier enjoys the advantages of multiple feature selection methods. The existing methods are least focused on ensembling. Hence, by the use of ensemble technique, the performance of feature selection can be improved.

Table 1: Comparison of recent works related to imbalanced classification.

Reference	Technique	Methodology	Comments
[23]	Extreme gradient boosting	Time-, frequency-, and spatial-based features were extracted by the proposed algorithm. Random forest is used for classification.	Correlation in time-based features can be improved. Embedded FS can be incorporated.
[24]	Orthogonal least squares	The authors have improved the speed of fetching the best features using orthogonal least squares. They have compared mutual information and other embedded methods.	Multiple correlation coefficient and the canonical correlation coefficient can be improved when feature generation and instance generation methods are used.
[25]	Centroid mutation-based search	A set of features which can represent a strong convergence to a set of classes is identified. This increases the position of classification margin and reduces the error.	The noisy features can be identified and removed before finding the strong convergence.
[26]	Balanced pointwise mutual information	A deep learning model is employed in Twitter text classification. Special characters like emoji are used as features to classify tweets.	Spam detection can be implemented to increase the accuracy.
[27]	Term weighting	Most of the feature selection methods just use frequency. The authors used category information as additional metric to select features for classification.	Semantics information can degrade the performance of the classification.

3. Relevant-Based Feature Ranking Algorithms

Feature selection is one of the important steps in text classification. The existing problem in ranking features is lack of identification of dependence. A good feature is identified by the following characteristics:

(i) A feature present in only one class is uniqueness, and it helps to identify the class correctly

(ii) A feature present in all the classes is not a good sign to identify a class

(iii) A feature is absent in one or more classes is also uniqueness, and it helps in negativity test

Consider a sample dataset as described in Table 2. There are two classes; one class is representing the topic astronomy, and other class is representing the topic society. Let us take the feature "planet" which is a unique feature in the topic 1; similarly, the feature "marriage" is a unique feature for the topic 2. The words "people" and "life" are present in both the topics. The ACC2 ratings are displayed in the last column; it is noted that for the unique feature "planet" and the nonsignificant feature "life" have the same rating, which is not a good sign for the classification. Hence, the rating methodology should be optimized to select the rich features.

The proposed feature selection algorithm takes this ranking problem in consideration and is aimed at assigning a rank based on its relevance towards the target class. If the feature represents the class fully, then high weight is given; similarly, when the feature is present in almost all the classes, then it is less likely that the proposed algorithm will pick this particular feature. The RBFR algorithm works in the following steps:

(1) Rank the features based on TPR-FPR

(2) Within the list, remove the features with low FPR

(3) Merge three filter-based FS algorithm selected features

(4) Rank the features based on class unique weights

Table 2: Problems in feature selection.

Feature	Class 1	Class 2	TRP	FPR	Accuracy measure (ACC2)
Planet	5	0	0.5	0	0.5
People	1	2	0.1	0.2	0.1
Life	9	4	0.9	0.4	0.5
Marriage	0	9	0	0.9	0.9

The feature ranks are given based on four metrics known as true positive (TP), true negative (TN), false positive (FP), and false negative (FN), which are defined as follows:

(i) TP: if a feature is present in the positive class

(ii) TN: if a feature is absent in the positive class

(iii) FP: if the feature is present in the negative class

(iv) FN: if the feature is absent in the negative class

The rich features for each class are determined by the ACC2 (TPR-FPR) [35], but there are high chances that the negative features are also selected alone with the rich features. Hence, a second level filtration on the basis of FPR could remove the weakly represented features.

3.1. Feature Selection Methods. To increase the rate of representation, three popular feature selection methods, namely, information gain, chi-square, and Pearson correlation, are

```
Input: F= set of features in the text corpus
Output: S – top N rich features
Begin:
For each f in F
    TPR score = $TP/TP + FN$
    FPR score = $FP/TN + FP$
L = {top $k_1$ features with high TPR-FPR score}
For each f in L
    If FPR(f)<TH then
        Remove f from L
    F1 = top N features from IG = $-\sum_{i=1}^{m} P(c_i)\ \log\ P(c_i) + P(t)\sum_{i=1}^{m} P(c_i|t)\ \log\ P(c_i|t) + P(\bar{t})\sum_{i=1}^{m} P(c_i|\bar{t})\ \log\ P(c_i|\bar{t})$
    F2 = top N features from CHI = $t(tp, (tp + fp) + ve_{prob} + t(fn, (fn + tn) + ve_{prob}) + t(fp, (tp + fp) - ve_{prob)} + t(tn, fn + tn) - ve_{prob})$
    F3 = top N features from Pearson Correlation
    Common Features = $(F1 \cap F2) \cup (F1 \cap F3) \cup (F2 \cap F3)$
Return $L \cup Common\ Features$
```

Algorithm 1: RBFR.

used to extract features. If a feature is selected by at least two of the feature selection methods, then that feature is also selected as per equation (4) for classification.

$$F = (F1 \cap F2) \cup (F1 \cap F3) \cup (F2 \cap F3). \tag{4}$$

$F1$, $F2$, and $F3$ in equation (4) represent the features selected by information gain, chi-square, and Pearson correlation, respectively. The details of the feature selection algorithms are briefed in the following subsections.

3.2. Class Unique Features. A feature is important based on how it represents the class. If a feature is present in only one class, then the feature is very important because it is very unique to a class. Similarly, if a feature is present across many classes, then it is very less important. After the second level of filtrations, a unique weight is calculated for each feature. This weight is based on the occurrence of a feature across various classes. Consider Table 3 which displays feature wise and class wise frequency, where $F_{i,j}$ represents the frequency of feature i in the class j. The first step is to remove the less class wise frequent term as per the condition in

$$\sum_i \sum_j F_{i,j} > \frac{\sigma}{n}. \tag{5}$$

The average of all frequency count is calculated, and the first step is to remove all the entries which have the frequency less than the average frequency. Then, an inverse class frequency is calculated to find out whether a feature is common or rare. A term which is very important is then filtered using a threshold value as described in equation (6), where $|C|$ is the total number of classes in the classification. $F(c)$ represents the number of classes the feature f represent.

$$\text{Threshold}(f) = \text{TF} * \log\left(\frac{|C|}{|F(\text{c})|}\right). \tag{6}$$

Table 3: Feature weights.

Feature	Class 1	Class 2	Class 3	Class 4
1	$F_{1,1}$	$F_{1,2}$	$F_{1,3}$	$F_{1,4}$
2	$F_{2,1}$	$F_{2,2}$	$F_{2,3}$	$F_{2,4}$
3	$F_{3,1}$	$F_{3,2}$	$F_{3,3}$	$F_{3,4}$

Table 4: Dataset description.

#	Dataset name	Number of documents	Number of features
1	Reuters [36]	1504	2886
2	WAP [37]	1560	6852
3	20 newsgroup [38]	18828	17425

3.3. Machine Learning Models. The proposed feature selection algorithm is tested using five machine learning models which are briefed in the following subsections.

3.3.1. k Nearest Neighbor. kNN is the machine learning models that finds distances between each instance. When a new sample or instance needs to be classified, the kNN finds the k closest neighbors from the instance, and the target class is found by majority voting. Some statistical methods are used to fix the value of K before starting the classification. It is better to fix the value of K as odd number. kNN is called as lazy classifier because it does nothing in the training phase; the distance calculation and the majority voting are done only in the classification phase.

3.3.2. Naïve Bayes. One of the most used classifiers in the field of text classification is Naïve Bayes. This model works with the probability concept of Bayes theorem. NB groups the instances based on similarity and determines the class of the new sample based on how much it is related with each class.

3.3.3. Support Vector Machines. Support vector machines are the most used classifier in the text classification domain.

Table 5: Performance in Reuters.

Feature selection	kNN			NB			SVM			RF			LR		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
CHI	52	55	56	83	60	70	11	67	19	81	96	88	89	78	83
ACC2	56	79	66	77	89	86	60	90	72	83	96	89	73	93	82
NDM	71	80	75	86	90	88	78	96	86	83	77	80	75	86	80
IF	76	67	71	81	98	89	97	85	91	95	87	91	94	84	89
GI	89	88	88	85	74	79	93	81	87	62	94	75	52	71	60
RBFR	64	95	76	93	97	95	94	84	89	95	92	93	91	89	90

Table 6: Performance in WAP.

Feature selection	kNN			NB			SVM			RF			LR		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
ACC2	38	32	35	88	86	87	54	67	60	82	99	90	58	97	73
CHI	83	91	87	87	94	90	94	75	83	94	84	89	7	14	9
NDM	60	88	71	65	94	77	25	22	23	93	47	62	75	87	81
GI	98	74	84	90	84	87	83	93	88	91	98	94	83	52	64
IF	80	88	84	92	87	89	88	96	92	93	97	95	91	81	86
RBFR	91	95	93	94	95	95	86	89	87	97	98	97	68	71	69

Table 7: Performance in 20 newsgroup.

Feature selection	kNN			NB			SVM			RF			LR		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
ACC2	98	64	77	87	94	90	67	70	68	92	91	91	82	86	84
CHI	93	79	85	92	90	91	79	79	79	95	94	94	77	72	74
NDM	97	78	86	92	60	73	92	96	94	74	79	79	59	57	58
GI	81	91	86	99	95	97	91	98	94	95	92	92	80	40	53
IF	90	86	88	84	93	88	86	86	86	96	96	96	95	58	72
RBFR	98	90	94	91	91	91	98	97	97	95	96	96	44	64	52

SVM can classify both linear as well as nonlinear data. A support vector is an end point in each class. The SVM model fixes a linearly separatable margin between the class; this margin is used to classify the instances.

3.3.4. Random Forest. RF is an ensemble-based classifier. The RF uses multiple decision tree. The number of DT is fixed before the start of classification. Each decision tree receives unique set of input and trained separately. Then, the output of each DT is used in majority voting to determine the final class.

3.3.5. Logistic Regression. LR is a special type of classifier that is used to classify linear data. LR constructs a margin which separates the classes. The new instances are assigned a class based on the position where it resides with respect to the margin.

4. Results and Discussion

We have used three benchmark datasets for evaluating our proposed feature selection algorithm. Table 3 contains the descriptions of all datasets.

4.1. Dataset Description. The three datasets contain different instances, number of classes, and number of features as shown in Table 4. We have taken random 2500 features from each dataset for our experiment.

4.2. Performance Evaluation. In order to test the performance of our proposed feature selection algorithm, we have used four standard metrics: accuracy, precision, recall, and F1-score. The formulas for calculating all the metrics are shown as follows:

$$\begin{aligned}\text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},\\ \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}},\\ \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{TN}},\\ \text{F1-Score} &= 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}.\end{aligned} \tag{7}$$

All the documents are preprocessed; stemming and stop word removal are done before the classification; also, a

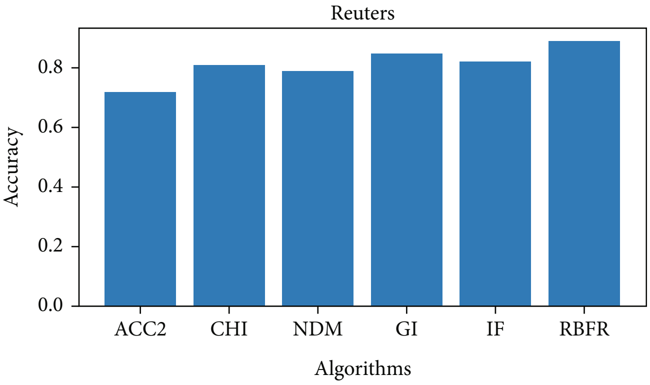


Figure 1: The accuracy comparison of kNN in Reuters dataset.

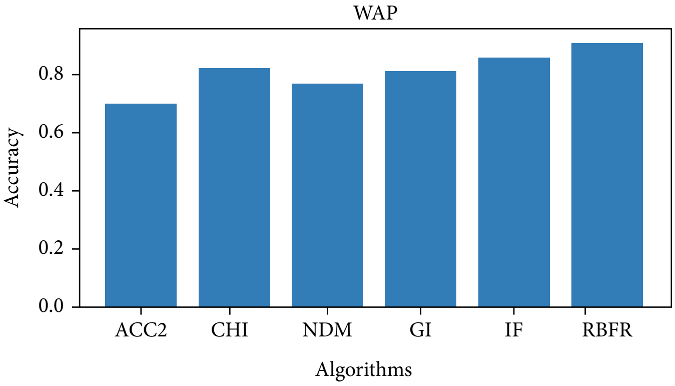


Figure 2: The accuracy comparison of kNN in WAP dataset.

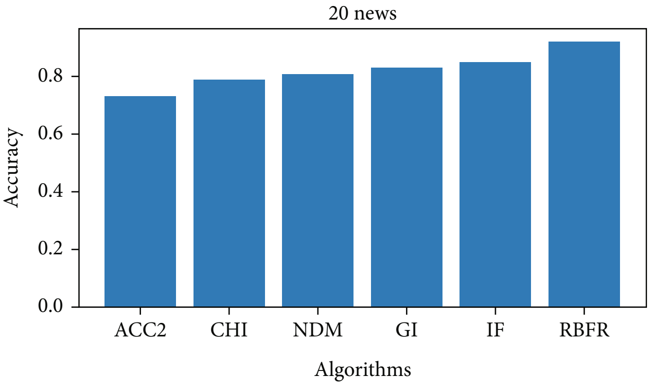
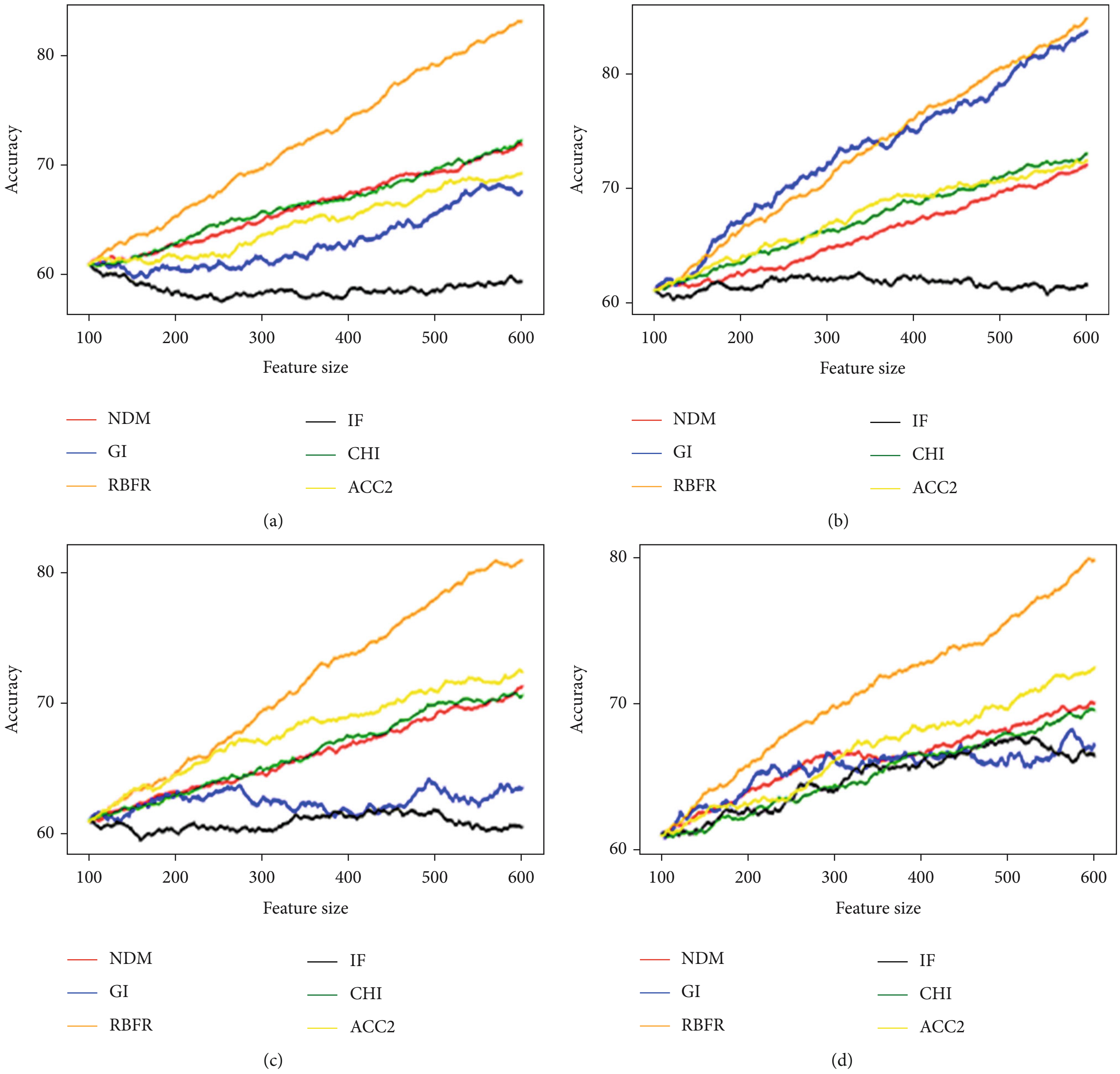


Figure 3: The accuracy comparison of kNN in 20 newsgroup dataset.

Table 8: Accuracy comparison.

Model	Two stage [39]	Noun based [40]	RBFR
LR	81.79%	74.91%	87.01%
kNN	85.94%	76.48%	89.6%
SVM	87.12%	81.44%	92.13%
NB	90.31%	87.8%	93.96%
RF	88.32%	88.91%	92.47%



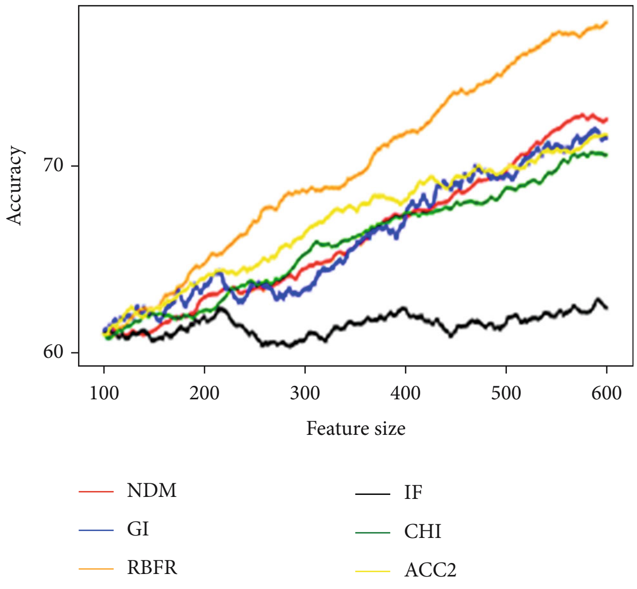
(a)

(b)

(c)

(d)

Figure 4: Continued.



(e)

Figure 4: The accuracy comparison when the number of features changes: (a) kNN classifier; (b) LR classifier; (c) NB classifier; (d) SVM classifier; (e) RF classifier.

second level of filtering is done by identifying the frequent words. The frequent words are features that are present in almost every document. The datasets are divided as per 10-fold validation.

The characteristics of the features that are selected by a feature selection algorithm can be analyzed to test the effectiveness of the feature selection algorithm. If unique features are selected and high rank is given to those features, then it is more likely that the performance of the classification will be good. Similarly, if irrelevant features are assigned higher ranks, then that will cause very poor performance in classification. The proposed feature selection method removes the high false rates thus provides a way to rank good feature. This is one of the reasons for the good performance of each classifier. Along with the ranking, the RBFR also considers top selected features from three well-known filter-based methods, and the common features present in them were selected. The precision, recall, and F1 comparison are shown in the Tables 5–7 for the datasets Reuters, WAP, and 20 newsgroups, respectively.

From the performance comparison tables, it is clear that the RBFR method identifies the rich features present in the corpus and ranks them higher than the irrelevant features. Precision is one of the good measures to judge a classification. It indicates the quality of positive predictions. The RBFR has higher precision in majority cases while compared with other feature selection methods.

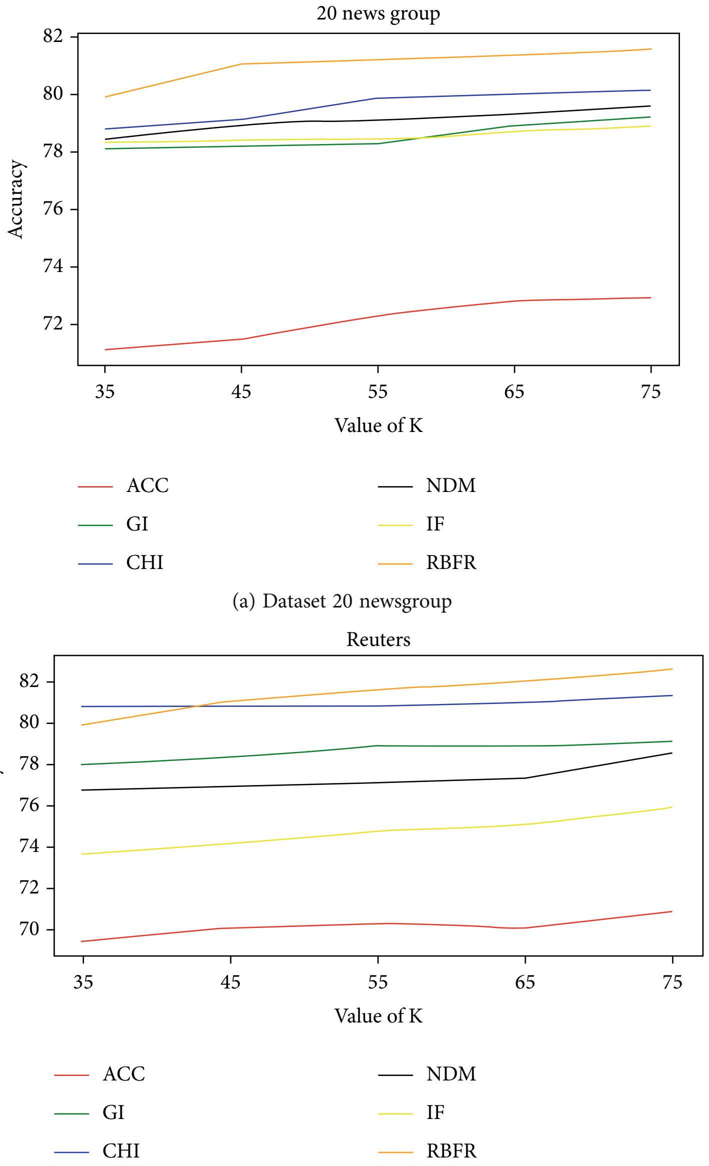
The ensemble of three filter-based feature selection increases the chance of selecting high rich features. As the selected features contain high level features, the classification using RBFR method is much higher than the classification done by other feature selection algorithms. Figures 1–3 display the accuracy of kNN in the three datasets. We have compared our proposed feature selection algorithm with other two works, and Table 8 shows the comparison.

The participation of multiple number of features in the process of classification is one of the important stages as it is not only responsible for increasing the efficiency of classification but also reduces the presence of simultaneous information redundancy. To solve the problems which affect the classification performance, the number of features should be selected optimally. If the feature size is very high, it increases the time of training rapidly; also if the size is too small, the accuracy becomes very low. Hence, the optimal number of features is determined by linearly increasing the number of features and stop when the performance degradation is observed.

In our experiment, we noticed that the optimal feature size is 600; after that, the accuracy of the classifiers seems to reduce. Among the classifiers, random forest seems to have increased accuracy even after 600; this is because the random forest can reduce the dimensionality by branching over the data. Up to 1400, the random forest classifier produces acceptable accuracy.

From Figure 4, it can be seen that among the existing feature selection methods, our proposed method outputs better performance in terms of accuracy, and SVM classifier produces the best accuracy when the number of features is 600. From the analysis, it can be found that as the number of features increases, there is a positive fluctuation in the classification performance. This is because, more sufficient knowledge can be derived in the training stage to improve the accuracy of the classification. Information duplication may arise when the number of features is increased too much; hence, an optimal count is preferred.

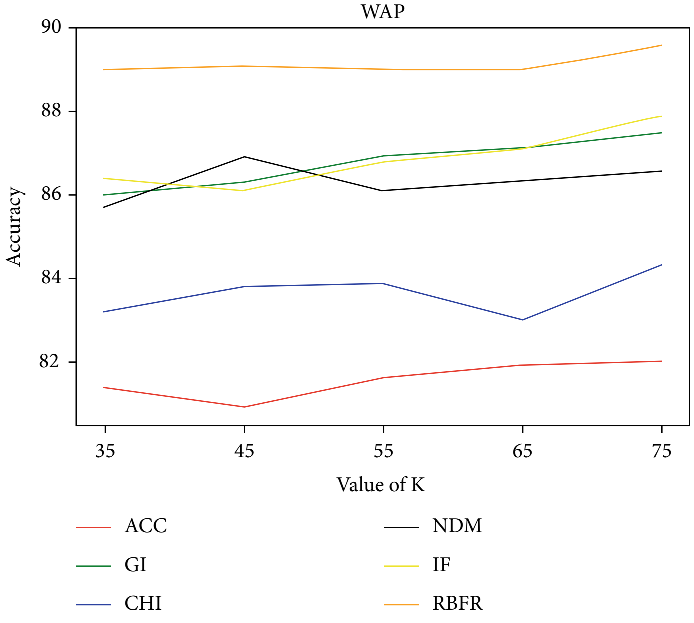
The number of neighbors plays a critical role in classification. From Figure 5, it can be observed that as the number of



(a) Dataset 20 newsgroup

(b) Reuters

Figure 5: Continued.



(c) WAP

Figure 5: The accuracy comparisons when different numbers of neighbors were considered in all three: (a) datasets 20 newsgroup; (b) Reuters; (c) WAP.

neighbor's increases, the performance also increases, but after 75, the classifier stabilizes. The proposed feature selection produces better results than the other feature selection methods because the removal of noise and redundant features.

5. Conclusions

Feature selection is one of the important stages in improving the performance of text classification. The existing feature selection methods can identify rich features present in the text corpus, but still lots of irrelevant features are also selected which degrades the performance of the text classification. In this work, we propose a ranking-based feature selection model which can identify and eliminate the irrelevant features from the selection set. We have implemented the proposed feature selection model in three datasets and compared with five existing filter-based feature selection methods, namely, ACC2, NDM, CHI, GI, and IG. The machine learning models used for classification were kNN, SVM, NB, LR, and RF. The experiment result shows that NB outperforms the classification task with 93.96% accuracy. In future work, we aim to rank the features based on its semantics and implement deep learning-based classification.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

References

[1] A. A. A. Ali and S. Mallaiah, “Intelligent handwritten recognition using hybrid CNN architectures based-SVM classifier with dropout,” *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 6, pp. 3294–3300, 2022.

[2] K. M. El Hindi, R. R. Aljulaidan, and H. AlSalman, “Lazy fine-tuning algorithms for naive Bayesian text classification,” *Applied Soft Computing*, vol. 96, article 106652, 2020.

[3] Z. Chen, L. J. Zhou, X. Da Li, J. N. Zhang, and W. J. Huo, “The Lao text classification method based on KNN,” *Procedia Computer Science*, vol. 166, pp. 523–528, 2020.

[4] A. Mohammed and R. Kora, *An Effective Ensemble Deep Learning Framework for Text Classification*, Journal of King Saud University-Computer and Information Sciences, 2021.

[5] T. H. J. Hidayat, Y. Ruldeviyani, A. R. Aditama, G. R. Madya, A. W. Nugraha, and M. W. Adisaputra, “Sentiment analysis of Twitter data related to Rinca Island development using Doc2-Vec and SVM and logistic regression as classifier,” *Procedia Computer Science*, vol. 197, pp. 660–667, 2022.

[6] Y. Li, X. Nie, and R. Huang, “Web spam classification method based on deep belief networks,” *Expert Systems with Applications*, vol. 96, pp. 261–270, 2018.

[7] K. Rakshitha, H. M. Ramalingam, M. Pavithra, H. D. Advi, and M. Hegde, “Sentimental analysis of Indian regional languages on social media,” *Global Transitions Proceedings*, vol. 2, no. 2, pp. 414–420, 2021.

[8] K. E. Daouadi, R. Z. Rebaï, and I. Amous, “Optimizing semantic deep forest for tweet topic classification,” *Information Systems*, vol. 101, article 101801, 2021.

[9] A. Palanivinayagam and S. Nagarajan, “An optimized iterative clustering framework for recognizing speech,” *International Journal of Speech Technology*, vol. 23, no. 4, pp. 767–777, 2020.

[10] E. Zhu, J. Zhang, J. Yan, K. Chen, and C. Gao, *N-gram MalGAN: evading machine learning detection via feature n-gram*, Digital Communications and Networks, 2021.

[11] D. Jain and V. Singh, "Feature selection and classification systems for chronic disease prediction: a review," *Egyptian Informatics Journal*, vol. 19, no. 3, pp. 179–189, 2018.

[12] L. Sun, T. Wang, W. Ding, J. Xu, and Y. Lin, "Feature selection using Fisher score and multilabel neighborhood rough sets for multilabel classification," *Information Sciences*, vol. 578, pp. 887–912, 2021.

[13] D. Paul, A. Jain, S. Saha, and J. Mathew, "Multi-objective PSO based online feature selection for multi-label classification," *Knowledge-Based Systems*, vol. 222, article 106966, 2021.

[14] F. BenSaid and A. M. Alimi, "Online feature selection system for big data classification based on multi- objective automated negotiation," *Pattern Recognition*, vol. 110, article 107629, 2021.

[15] A. Hashemi, M. B. Dowlatshahi, and H. Nezamabadi-pour, "An efficient Pareto-based feature selection algorithm for multi-label classification," *Information Sciences*, vol. 581, pp. 428–447, 2021.

[16] M. Selvapriya and G. M. Priscilla, "Integrated feature selection (IFS) algorithm and enhanced weight based convolutional neural network (EWCNN) for social emotion classification," *Materials Today: Proceedings*, 2021.

[17] E. O. Omuya, G. O. Okeyo, and M. W. Kimwele, "Feature selection for classification using principal component analysis and information gain," *Expert Systems with Applications*, vol. 174, article 114765, 2021.

[18] J. Chong, P. Tjurin, M. Niemelä, T. Jämsä, and V. Farrahi, "Machine-learning models for activity class prediction: a comparative study of feature selection and classification algorithms," *Gait & Posture*, vol. 89, pp. 45–53, 2021.

[19] C. F. Tsai, K. L. Sue, Y. H. Hu, and A. Chiu, "Combining feature selection, instance selection, and ensemble classification techniques for improved financial distress prediction," *Journal of Business Research*, vol. 130, pp. 200–209, 2021.

[20] W. Qian, C. Xiong, and Y. Wang, "A ranking-based feature selection for multi-label classification with fuzzy relative discernibility," *Applied Soft Computing*, vol. 102, article 106995, 2021.

[21] Z. Ruijie, X. Ying, J. Shuaichen, and L. Yonghe, "Patent text modeling strategy and its classification based on structural features," *World Patent Information*, vol. 67, article 102084, 2021.

[22] J. Dai and J. Chen, "Feature selection via normative fuzzy information weight with application into tumor classification," *Applied Soft Computing*, vol. 92, article 106299, 2020.

[23] J. Ma and X. Gao, "A filter-based feature construction and feature selection approach for classification using genetic programming," *Knowledge-Based Systems*, vol. 196, article 105806, 2020.

[24] T. Thenmozhi and R. Helen, "Feature selection using extreme gradient boosting Bayesian optimization to upgrade the classification performance of motor imagery signals for BCI," *Journal of Neuroscience Methods*, vol. 366, article 109425, 2022.

[25] S. Zhang and Z. Q. Lang, "Orthogonal least squares based fast feature selection for linear classification," *Pattern Recognition*, vol. 123, article 108419, 2022.

[26] E. H. Houssein, E. Saber, A. A. Ali, and Y. M. Wazery, "Centroid mutation-based search and rescue optimization algorithm for feature selection and classification," *Expert Systems with Applications*, vol. 191, article 116235, 2022.

[27] Z. Ahanin and M. A. Ismail, "A multi-label emoji classification method using balanced pointwise mutual information-based feature selection," *Computer Speech & Language*, vol. 73, article 101330, 2022.

[28] F. Shen, X. Zhang, R. Wang, D. Lan, and W. Zhou, "Sequential optimization three-way decision model with information gain for credit default risk evaluation," *International Journal of Forecasting*, vol. 38, no. 3, pp. 1116–1128, 2022.

[29] P. Jagadeesan, K. Raman, and A. K. Tangirala, "A new index for information gain in the Bayesian framework*," *IFAC-Papers OnLine*, vol. 53, no. 1, pp. 634–639, 2020.

[30] N. Peker and C. Kubat, "Application of chi-square discretization algorithms to ensemble classification methods," *Expert Systems with Applications*, vol. 185, article 115540, 2021.

[31] S. Bahassine, A. Madani, and M. Kissi, "An improved Chi-sqaure feature selection for Arabic text classification using decision tree," in *2016 11th International Conference on Intelligent Systems: Theories and Applications (SITA)*, pp. 1–5, IEEE, 2016.

[32] B. Kalaiselvi and M. Thangamani, "An efficient Pearson correlation based improved random forest classification for protein structure prediction techniques," *Measurement*, vol. 162, article 107885, 2020.

[33] M. Baak, R. Koopman, H. Snoek, and S. Klous, "A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics," *Computational Statistics & Data Analysis*, vol. 152, article 107043, 2020.

[34] P. Ashokkumar and S. Don, *Link-based clustering algorithm for clustering web documents*, ASTM International, 2018.

[35] A. Rehman, K. Javed, and H. A. Babri, "Feature selection based on a normalized difference measure for text classification," *Information Processing & Management*, vol. 53, no. 2, pp. 473–489, 2017.

[36] "Reuters dataset," https://archive.ics.uci.edu/ml/datasets/reuters21578+text+categorization+collection.

[37] "WAP dataset," http://glaros.dtc.umn.edu/.

[38] "20 news group dataset," http://qwone.com/~jason/20Newsgroups/.

[39] P. Ashokkumar, S. G. Shankar, G. Srivastava, P. K. Maddikunta, and T. R. Gadekallu, "A two-stage text feature selection algorithm for improving text classification," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 3, pp. 1–19, 2021.

[40] G. Siva Shankar, P. Ashokkumar, R. Vinayakumar, U. Ghosh, W. Mansoor, and W. S. Alnumay, "An embedded-based weighted feature selection algorithm for classifying web document," *Wireless Communications and Mobile Computing*, vol. 2020, 10 pages, 2020.