Research Article GU J Sci, Part A, 10(4): 472-486 (2023) 10.54287/gujsa.1379024

Gazi University
Journal of Science
PART A: ENGINEERING AND INNOVATION
http://dergipark.org.tr/gujsa

A New Feature Selection Metric Based on Rough Sets and Information Gain in Text Classification

Rasim ÇEKİK[1] Mahmut KAYA[2*]

[1] Department of Computer Engineering, Şırnak University, Şırnak, Türkiye
[2] Department of Computer Engineering, Siirt University, Siirt, Türkiye

Keywords

Feature Selection
Text Classification
Rough Set
Dimensiomality Reduction

Abstract

In text classification, taking words in text documents as features creates a very high dimensional feature space. This is known as the high dimensionality problem in text classification. The most common and effective way to solve this problem is to select an ideal subset of features using a feature selection approach. In this paper, a new feature selection approach called Rough Information Gain (RIG) is presented as a solution to the high dimensionality problem. Rough Information Gain extracts hidden and meaningful patterns in text data with the help of Rough Sets and computes a score value based on these patterns. The proposed approach utilizes the selection strategy of the Information Gain Selection (IG) approach when pattern extraction is completely uncertain. To demonstrate the performance of the Rough Information Gain in the experimental studies, the Micro-F1 success metric is used to compare with Information Gain Selection (IG), Chi-Square (CHI2), Gini Coefficient (GI), Discriminative Feature Selector (DFS) approaches. The proposed Rough Information Gain approach outperforms the other methods in terms of performance, according to the results.

Cite

Cekik, R., & Kaya, M. (2023). A New Feature Selection Metric Based on Rough Sets and Information Gain in Text Classification. *GU J Sci, Part A, 10*(4), 472-486. doi:10.54287/gujsa.1379024

Author ID (ORCID Number)

0000-0002-7820-413X Rasim ÇEKİK
0000-0002-7846-1769 Mahmut KAYA

Article Process

Submission Date 20.10.2023
Revision Date 08.11.2023
Accepted Date 17.11.2023
Published Date 12.12.2023

1. INTRODUCTION

Several applications, such as e-commerce, social networks, location-based services, and information collecting and distribution centers, have emerged as a result of technology's rapid progress. This has resulted in the creation of big data in the digital environment. Most of this data consists of text documents. Examples of text documents include news published on news sites, tweets and posts on social media, and comments about products on e-commerce sites. However, in information gathering and distribution centers, people prefer to see only the web pages (or documents) that interest them and do not want to see irrelevant ones. Therefore, text data needs to be categorized and indexed according to their content. This allows web pages to be easily searched. But since text data has a large volume, it is impossible to process it manually. To overcome this challenge, automatic document classification methods have been proposed. This process is called text classification or categorization. One of the main problems in the field of text classification is that using one or more of each word in the document as a feature lead to high dimensional data problems. To provide a solution to this problem, feature selection methods aim to select the optimal subset in the entire feature set space. This optimal set consists of highly discriminative terms and is expected to be the subset that best represents the entire feature set. The basic purpose of feature selection methods is to select the best, most optimal set of features. Selection techniques are one of the methods that increase the performance and success of decision systems such as machine learning and reduce the execution time (Kaya et al., 2013; Kaya & Bilge, 2016; Şenol, 2023). Therefore, the success of the selection technique is an important factor in problem solving.

*Corresponding Author, e-mail: mahmutkaya@siirt.edu.tr

According to their working principles, feature selection techniques can be grouped under three headings: filter, embedding and wrapper. Filter methods calculate a score for each feature separately and work by selecting n features with the highest scores. Embedded techniques integrate the feature selection process directly into the classification algorithm and this approach is often used in certain learning algorithms. Wrapper techniques try to select the optimal subset of features using a classifier. In the literature, there are several feature selection approaches related to these different techniques. For example, improved gini index (IGI) (Shang et al., 2007), distinguishing feature selector (DFS) (Uysal & Gunal, 2012), global information gain (GIG) (Shang et al., 2013), class discriminating measure (CDM) (Chen et al., 2009), multivariate relative discrimination criterion (MRDC) (Labani et al., 2018), Fisher's discriminant ratio (Wang et al., 2009), normalized difference measure (NDM) (Rehman et al., 2017), max-min ratio (MMR) (Rehman et al., 2018) and proportional rough feature selector (PRFS) (Cekik & Uysal, 2020) are known correlation-based methods. Frequency-based, evolutionary algorithm-based and other theory-based techniques are frequently used in the literature for this purpose.

In this study, a new filter feature selector method called Rough Information Gain (RIG) based on correlation-based Rough Set Theory and Information Gain is proposed. The proposed method works by using different structures according to the three states "certain", "roughly certain" and "uncertain" in the text document information system. With the help of Rough Sets, the status of the text documents (certain, roughly certain and uncertain) is determined. If the state is uncertain, then Information Gain comes into play and a score value is calculated for the state. In other cases, Rough Sets and Information Gain work together to produce a score, and finally a score value for each feature.

2. RELATED WORKS

Feature selection is the procedure of selecting the most representative, highly discriminative features from a set of accessible features using a feature selection algorithm. This concept has been addressed in the literature with many studies investigating different feature selection approaches. For example, there are traditional methods such as Information Gain (IG), Gain Ratio (GR), Gini Index (GI), Chi2, Mutual Information (MI) (Sharmin et al., 2019) as well as recently proposed approaches such as DFS, NDM, MMR and MRDC. Many of these methods are widely used in applications such as text classification (Cekik & Uysal, 2022). The IG approach is commonly used, particularly in data and text mining. This method is centered on Shannon's information theory and thermodynamic concepts. However, if there are many different values that an attribute can take, the IG method may select it as an attribute that is easily memorized by the system. GR computes the gain ratio by dividing the discrimination information for each feature by the information gain to overcome this problem. GI is an alternative feature selection approach to IG and GR and does not use the entropy value. The Chi2 method statistically evaluates the chi-square values of all features according to their class. MI is another approach for calculating the interdependence of a feature and a class label. However, it can tend to prioritize rare features and can be sensitive to errors in probability estimation. The DP method is a widely used approach in the field of information retrieval to identify influential words and has been adapted to feature selection problems (Ogura et al., 2009). These methods offer different approaches to the feature selection process and it is important to decide which method will work better depending on the context of the problem to be applied.

DFS is one of the recently proposed novel approaches for feature selection, in addition to the traditional techniques. DFS produces a value between 0.5 and 1.0 according to the feature importance of each term. One of the recent studies, RDC, was presented by Rehman et al. (2015) as a new feature selection approach. This method takes into account the document frequencies of each term when identifying highly discriminative terms. The MRDC method was developed by Labani et al. (2018). T The fitness value for each term is calculated using Pearson correlation, and the selected term set is evaluated using a supervised learning algorithm. There is also a new text classification method called NDM proposed by Rehman et al. (2017). This approach is designed for text classification using a balanced accuracy measure. However, considering that it is not effective on imbalanced datasets with high sparsity, a new approach called MMR is proposed by the same authors. Finally, Wang and Hong (2019) proposed HRFS, a Hebb-rule feature selection model for text classification. This model considers the class and terms as neurons and focuses on selecting terms with high discriminative power. Cekik and Uysal (2022) introduces a new feature selection approach called the XY method, which effectively operates on short texts. This approach works by computing the distance between

two points in a two-dimensional plane and the XY line. This point's position is determined by the number of documents in which a phrase appears in one class versus the number of documents in which it appears in another class.

In the literature, apart from filter techniques for text classification, methods such as Linear Forward Search (LFS) (Gutlein et al., 2009), Span-Bound and RW-Bound (Weston et al., 2001) are examples of wrapper approaches, while methods such as EGA (Ghareb et al., 2016), FSS (Bermejo et al., 2012), HybridBest and HybridGreedy (Chou et al., 2010) are examples of embedded approaches. It should be noted that in the field of text classification, filter approaches are represented by more studies than other methods. This perspective is the result of the fact that filter approaches can work faster and more efficiently on high-dimensional data. Consequently, it is a fact that filter-based feature selection methods still need better and more efficient solutions. The main motivation of this work is to present new filter-based feature selection approaches that can work effectively in the field of text classification.

3. PRELIMINARIES AND BACKGROUND

3.1. Rough Set Theory

Rough Set Theory (RST) (Pawlak, 1998; Zhang et al., 2016) is a mathematical approach proposed by Pawlak (1998) that aims to make efficient inferences on incomplete and inconsistent data. This theory offers a structure that can handle verified logic, inconsistent data and imprecise latent inferences, and avoids strict limitations such as fuzzy sets. It uses both fuzzy and rough set structures to organize incomplete, inadequate and uncertain information in terms of data analysis. In rough set theory, data is stored in the form of a table containing attributes and conditional attributes and adopts the concept of equality class to divide the training data according to certain criteria. In the learning process, two types of partitions, low approximation and high approximation, are created to obtain exact and probabilistic rules. High approximation refers to elements that are unequivocally part of the set, while low approximation represents elements that are likely to belong to the set. Rough set theory plays an important role in data analysis and learning processes by handling incomplete and inconsistent data.

The rough set approach is based on two concepts, low and high approximation.

- Elements that are certain to belong to the set,
- Elements that are expected to be part of the set

Explanations of the basic concepts of rough sets are briefly stated and italicized. An example of a decision table is given in Table 1. The decision table is known as the table or information system that holds the data in Rough Sets (RS). $S = (U, A, C)$ denotes a decision table or information system, where $U = \{x_1, x_1, \dots, x_n\}$ is the universal set of objects, A is a conditional attribute set and C is a decision attribute set. For any conditional attribute subset $T \subseteq A$, the T-indistinguishability relation, denoted $IND(T)$, is defined as follows:

$$IND(T) = \{(x_i, x_j) \in U^2 | \forall a \in T, a(x_i) = a(x_j)\} \quad (1)$$

Where, the equivalence classes of the T-indistinguishability relation are expressed as $[x]_T$.

The lower and upper set approximations illustrate two essential rough set notions. The notation for T-lower and T-upper approximations of the set X over any subset $X \subseteq U$ of objects and a given subset $T \subseteq A$ of attributes is $\underline{T}X$ and $\bar{T}X$, respectively. They are also defined as follows:

$$\underline{T}X = \{x \mid [x]_T \subseteq X\}, \quad (2)$$

$$\bar{T}X = \{x \mid [x]_T \cap X \neq \emptyset\} \quad (3)$$

The pair ($\underline{T}X, \bar{T}X$), known as a rough set approximation concept, indicates whether or not a set may be roughly determined. This pair also determines the known regions in rough sets. If an object $x \in \underline{T}X$, it is known that it belongs to set X. But if $x \in \bar{T}X$, it may belong to set X, i.e. it is not certain that it belongs. For example, if $T = \{t_1, t_4\}$ and $X = \{x_1, x_2, x_5, x_7, x_8\}$ according to the decision table shown in Table 1 above, then $\underline{T}X = \{x_2, x_7\}$ and $\bar{T}X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$.

Table 1. *Example of a simple decision table*

$x \in U$	t_1	t_2	t_3	t_4	d
x_1	1	2	0	2	1
x_2	0	2	1	1	1
x_3	1	1	2	2	3
x_4	2	2	2	3	4
x_5	2	2	1	3	3
x_6	1	1	2	2	2
x_7	0	1	1	1	1
x_8	2	2	0	3	2

Where, the attributes x_2 and x_7 are certain to belong to the set X, while the other attributes may belong, i.e. their belonging is not certain. The definition of Positive ($POS_T X$), Negative ($NEG_T X$) and Boundary ($BND_T X$) regions defined based on the pair ($\underline{T}X, \bar{T}X$) is given below and a representative representation is given in Figure 1.

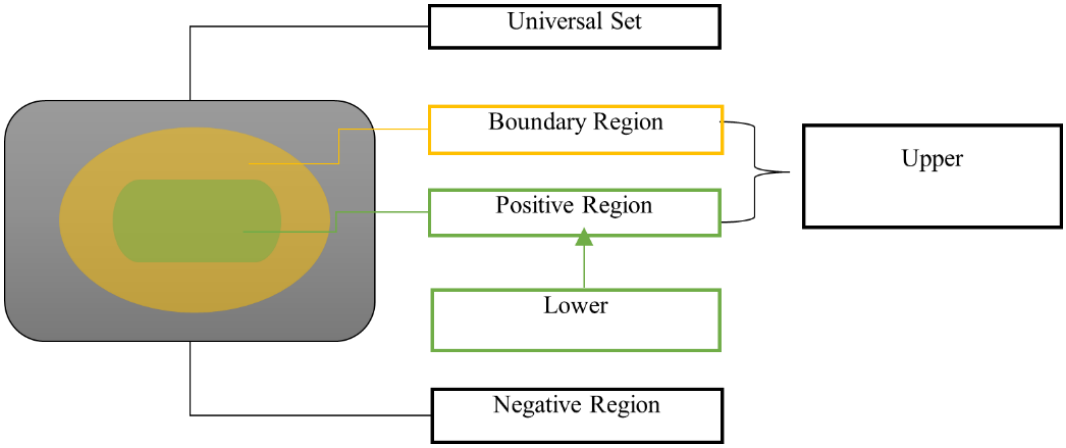


Figure 1. *The RST's regions and approximation sets*

Defining regions on sets X and T is as follows:

$$POS_T X = \underline{T}X \tag{4}$$

$$NEG_T X = U - \bar{T}X \tag{5}$$

$$BND_T X = \bar{T}X - \underline{T}X \tag{6}$$

According to RST, if the border region is not an empty set, then that set is said to be roughly determined. Otherwise, the set is said to be fully determined. For example, for sets $T = \{t_1, t_4\}$ and $X = \{x_1, x_2, x_5, x_7, x_8\}$, the regions are:

$$POS_T X = \{x_2, x_7\} \tag{7}$$

$$NEG_T X = \emptyset \tag{8}$$

$$BND_T X = \{x_1, x_3, x_4, x_5, x_6, x_8\} \tag{9}$$

The accuracy of the rough set, as expressed by Pawlak (1998), can be formulated with the following formula:

$$\lambda_T(X) = \frac{|\underline{T}(X)|}{|\overline{T}(X)|} \tag{10}$$

This is the accuracy of the rough set representation of set X, denoted as $\lambda_T(X)$. Here, $0 \leq \lambda_T(X) \leq 1$, and it represents the ratio of the number of objects that can be positively placed in set X to the number of objects that can possibly be placed in set X. This provides a measure of how closely the rough set approximates the target set. Clearly, when the upper and lower approximations are equal (i.e., the boundary region is empty), then $\lambda_T(X) = 1$, and the approximation is perfect. At the other extreme, regardless of the size of the upper approximation, if the lower approximation is empty, the accuracy is zero.

3.2. Information Gain

Information Gain (IG) (Yang & Pedersen, 1997) is a statistical information used in data and text mining that assesses the relevance of a term or word in a given text. This metric works by evaluating the salience of a term within a document and the probability that the term belongs to a specific category. It is commonly defined as the inverse of entropy. The mathematical formulation behind Information Retrieval:

$$IG(t) = -\sum_{i=1}^{M} P(C_i) log P(C_i) + P(t) \sum_{i=1}^{M} P(C_i|t) log P(C_i|t) + P(\bar{t}) \sum_{i=1}^{M} P(C_i|\bar{t}) log P(C_i|\bar{t}) \tag{11}$$

Where M represents the number of classes and $P(C_i)$ is the probability of class C_i. $P(C_i|t)$ and $P(C_i|\bar{t})$ denote the conditional probabilities of having class C_i at the same time when term t is present and having class C_i at the same time when term t is absent, respectively. Similarly, $P(\bar{t})$ and $P(t)$ denote the probabilities of t terms passing and not passing.

3.3. Preprocessing

Preprocessing is a very important and critical step in text classification. The preprocessing stage is also known as a sequence of operations on text collections, such as data cleaning, finding semantic values of words, data normalization and data integrity. In this process or stage, the following operations are generally performed:

- Cleaning unwanted (noise) data (deletion or correction of data due to spelling errors, etc.)
- Removing unnecessary words (conjunctions, prepositions, pronouns, etc.)
- Finding semantic values of words (noun, verb, adverb, etc.)
- Removing punctuation marks
- Dividing the text into sections or words
- Making lowercase to uppercase conversions
- Disassembling words into their roots (removal of suffixes, if any, etc.).

In addition to data cleaning, the preprocessing stage also aims to ensure data integrity and normalization, and to bring the data into the appropriate format. It is commonly divided into 4 categories: tokenization, stop-word removal, lowercase conversion and stemming. In this study, these operations were applied respectively.

4. THE PROPOSED METHOD

Rough Information Gain (RIG) successfully reveals the dependency of the attributes in an information system on the decision attribute and the characteristic of each attribute to the decision attribute. In this study, we compute the characteristics of each attribute with the help of upper and lower set approaches in RIG. Upper and lower approaches can understand the following characteristics of a set.

- whether it can be roughly determined.
- can be determined with certainty.
- can be determined to be uncertain.

If a set is roughly defined, a lower approach is used to obtain documents that definitely belong to that set, while a upper approach is used to obtain documents that are likely to belong to that set. With the proposed method, document sets are obtained according to the value set of each attribute. These document sets are determined whether each of them is complete, roughly complete or uncertain by upper and lower approaches. If the set is exact or rough, the characteristic value of the feature is calculated from formula (11):

$$\lambda_R(X) = \frac{|\underline{R}(X)|}{|\overline{R}(X)|} \tag{12}$$

Where R is the feature subset and X is the set of documents to be identified. $\lambda_R(X)$ is the accuracy of the set X.

If it cannot be roughly or exactly determined (uncertain) then a weight is calculated with IG (see Equation 11). The sum of the feature characteristic value and the GI value gives the weight of this set (the set allocated according to the feature) or feature:

$$RIG(t) = \lambda_R(X) + G(t) \tag{13}$$

The features are ranked according to this weight value and a total of n features are selected. The specified features are used to prepare both training and test data. As a result, now that the training and test data are ready, all that remains is to classify them with a classifier. The working mechanism of the proposed method RIG is shown in Figure 2.

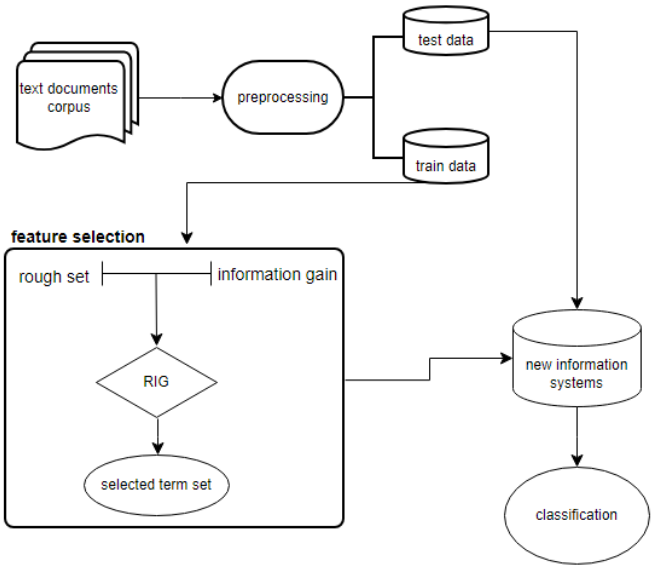


Figure 2. *The Proposed Method*

An Example:

Suppose we have 5 documents and each document belongs to either class C_1 or C_2 (C). The words in the documents are cat, dog and mouse. The following information system is given in Table 2.

Table 2. *A simple document collection*

	Terms			
Document Name	*cat*	*dog*	*mouse*	*Class Label (C)*
D_1	*1*	*0*	*0*	C_1
D_2	*1*	*1*	*0*	C_1
D_3	*1*	*2*	*1*	C_2
D_4	*1*	*1*	*0*	C_2
D_5	*1*	*0*	*1*	C_2

In the information system given in Table 2, the frequency of occurrence of each word in a document was ignored. The frequency of occurrence of the word in a document is ignored. According to the value set of each feature on Table 2, document clusters are created as follows:

- $If\ R = \{cat\}\ R = \{(D_1, D_2, D_3, D_4, D_5)\}$
- $If\ R = \{dog\}\ R = \{(D_1, D_5), (D_2, D_4), (D_3)\}$
- $If\ R = \{mouse\}\ R = \{(D_1, D_2, D_4), (D_3, D_5)\}$

Also document clusters by class;

Two separate clusters are identified as $X_1 = \{u\,|C(u) = C_1\} = \{D_1, D_2\}$ and $X_2 = \{u\,|C(u) = C_2\} = \{D_3, D_4, D_5\}$. In this case, scores can be calculated for each attribute. These are respectively:

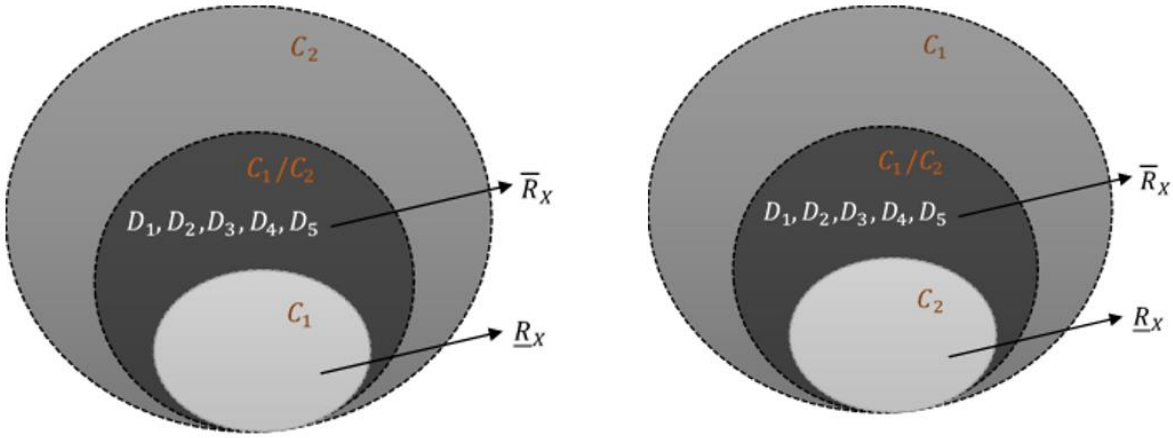
- ***for the term cat;***

Let calculate the lower and upper set approximations of the set X_1: $\underline{R}_X = \emptyset$ and $\overline{R}_X = \{D_1, D_2, D_3, D_4, D_5\}$. Moreover, the representative set representation is shown in Figure 3a. According to this, the set X_1 cannot be determined exactly or roughly. Similarly, the lower and upper set approximations are calculated for the set X_2: $\underline{R}_X = \emptyset$ and $\overline{R}_X = \{D_1, D_2, D_3, D_4, D_5\}$. A representative cluster representation is also shown in Figure 3b. The set X_2 cannot be determined exactly or roughly. Therefore, the weight value is calculated by applying IG for cluster $\overline{R}_X$ (for documents in the upper approach):

$$IG(\overline{R}_X) = 0.9710 - 0.9710 + 0 = 0 \text{ and } \lambda_R(X) = \frac{0}{5} = 0$$

Accordingly, the score of the cat term is as follows:

$$\text{RIG(cat)} = \lambda_R(X) + IG(\overline{R}_X) = 0$$



***(a)** According to set X_1* *(b) According to set X_2*

***Figure 3**. Upper and Lower sets according to cat term*

- ***for the term dog;***

Let calculate the lower and upper set approximations of the set X_1: $\underline{R}_X = \emptyset$ and $\overline{R}_X = \{D_1, D_2, D_4, D_5\}$. Accordingly, the set X_1 cannot be determined exactly or roughly. The representation is shown in Figure 4a.

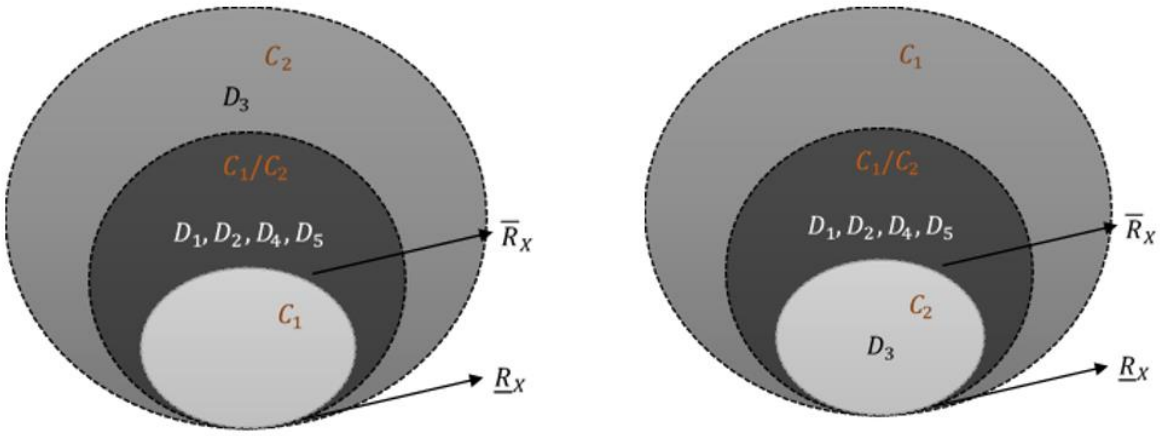
$$IG(\overline{R}_X) = 1 - 0.5 - 0.5 = 0$$

For the set X_2 :$\underline{R}_X = \{D_3\}$ and $\overline{R}_X = \{D_1, D_2, D_3, D_4, D_5\}$. Accordingly, the set X_1 cannot be determined exactly or roughly. A representative cluster representation is also shown in Figure 4b.

$$\lambda_R(X_2) = \frac{1}{5} = 0.2$$

Accordingly, the score of the dog term is as follows:

$$\text{RIG(dog)} = \lambda_R(X_2) + IG(\overline{R}_X) = 0.2$$



***(a)** X_1 According to set X_1* *(b) X_2 According to set X_2*

***Figure 4**. Upper and Lower sets according to dog term*

- ***for the term mouse***;

Let calculate the lower and upper set approximations of the set X_1: $\underline{R_X} = \emptyset$ and $\overline{R}_X = \{D_1, D_2, D_4\}$. Accordingly, the set X_1 cannot be determined exactly or roughly. Moreover, the representative cluster representation is shown in Figure 5a. Therefore, IG is applied for set X_1 to calculate the weight value:

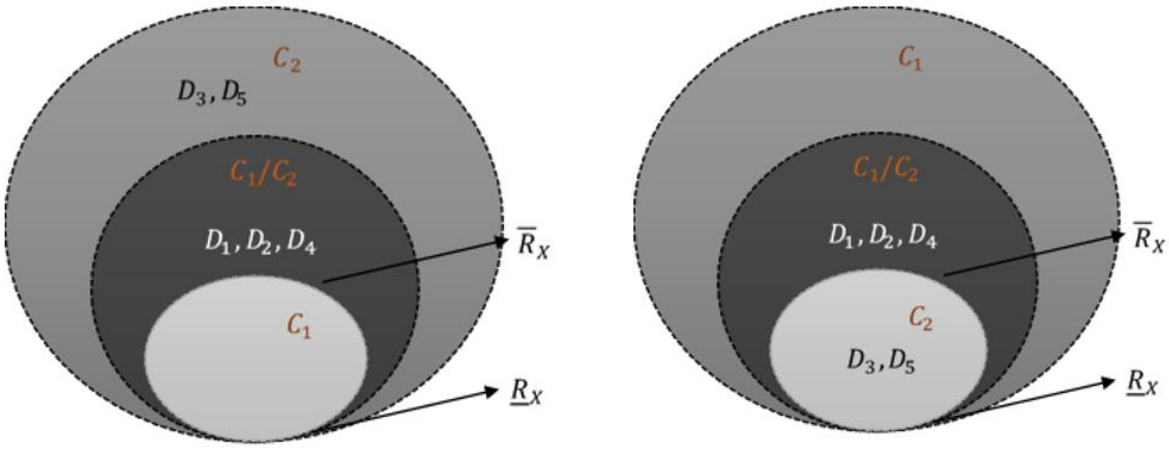
$$GI(X_1) = 0.9236 - 0.9236 = 0$$

For the set X_2: $\underline{R_X} = \{D_3, D_5\}$ and $\overline{R}_X = \{D_1, D_2, D_3, D_4, D_5\}$. Moreover, the representative set representation is shown in Figure 5b. This set is a roughly determined set. Therefore, the characteristic value is calculated for the set X_2 :

$$\lambda_R(X_2) = \frac{2}{5} = 0.4$$

Accordingly, the score of the dog term is as follows:

$$\text{RIG(mouse)} = \lambda_R(X_2) + IG(\overline{R}_X) = 0 + 0.4 = 0.4$$



(a) X_1 *According to set* X_1 ***(b)*** X_2 *According to set* X_2

Figure 5*. Upper and Lower sets according to mouse term*

As a result, cat = 0.0, dog = 0.2, mouse = 0.4 and according to top-3, the order: $mouse > dog > cat$ is obtained.

The RIG has made an effective choice in the feature selection process in the field of text classification by utilizing the effective extraction ability of rough sets when there is not enough data. Therefore, The RIG successfully reveals the dependence of attributes in an information system on the decision attribute and the dependence of each attribute's characteristic on the decision attribute. This feature distinguishes it from other existing methods.

5. EXPERIMENTAL RESULTS

5.1. Datasets

In the experiments, Eminem and Kat Peryy datasets from the English Short Message Service (SMS) (Nuruzzaman et al., 2011) dataset and Youtube Spam Collection (Alberto et al., 2015) datasets were used. The process performed on these datasets is to determine whether an SMS is spam or not and whether the comments in the Youtube Spam Collection dataset are spam or not. The first dataset is the dataset created by Nuruzzaman

et al. (2011), which contains 425 spam and 450 non-spam English SMS text messages. The other datasets, Eminem and Kat Perry, are a general set of comments collected for spam research. They account for two of the top five most watched videos during the collection period and consist of 548 and 300 real comments, respectively. Table 3 also provides general information about the datasets.

Table 3. *Datasets*

	Spam	*Non-Spam*
SMS	*440*	*425*
Eminem	*245*	*203*
Kat Peryy	*175*	*175*

5.2. Classifiers

Classification is the process of assigning previously unlabeled data to a labeled category in a dataset. There are many approaches presented in the literature for this process. In this study, the classification methods given in Table 4 below were used.

5.3. Measure of Success

In this section, the Micro-F1 metric is chosen to evaluate the performance of feature selection approaches. In the calculation of this metric, the harmonic mean of Precision and Recall is employed in order to summarize the performance of the classifier algorithm in a more balanced way. Precision is the number of true positive examples (TP) divided by the sum of the number of true positive examples (TP) and the number of false positive examples (FP).

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

Recall is defined as the ratio of true positive samples (TP) to the sum of true positive samples (TP) and false negative samples (FN).

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

Precision is especially important when the cost of false positive predictions is high, and Recall is important when the cost of predicting false negatives is high.

Accordingly, Micro-F1:

$$Mikro - F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{16}$$

Table 4. *General information about classifiers used in experimental studies*

Method	Mathematical Expression	Description
Support Vector Machines (Joachims, 1998) (SVM)	$w^T x + w_0 = 0.$ $J(w, w_0, \varepsilon) = \frac{1}{2}\|\|w\|\|^2 + K\sum_{i=1}^{N}\varepsilon_i$ $w^T x + w_0 \geq 1 - \varepsilon_i \quad eğer \ x_i \in c_1$ $w^T x + w_0 \leq -1 + \varepsilon_i \quad eğer \ x_i \in c_2$ $\varepsilon_i \geq 0.$ $w = \sum_{i=1}^{N}\lambda_i y_i x_i$	Based on margin maximization, SVM is one of the most efficient and well-known classifiers. There are linear and non-linear versions available. However, the linear version is given here. Its main goal is to obtain the highest possible margin. $2/\|\|w\|\|$ is the margin width, K denotes a user-defined constant, and ε represents the margin error.
k-Nearest Neighbors (Kowsari et al., 2019) (KNN)	$\left(\sum_{i=1}^{k}(\|x_i - y_i\|)^q\right)^{1/q}$	Using distance functions such as Euclidean to find the similarity of neighbors, KNN scores category candidates based on the class of k candidates by finding the k closest neighbors of a test data among all data in the training set. It is provided the Minkowski distance computation function. In this function, if q is $q = 1$, Manhattan and $q = 2$, Euclidean functions are obtained.
Decision Tree (Aggarwal & Zhai, 2012) (DT)	$H\left(\frac{p}{n+p}, \frac{n}{n+p}\right) = -(\frac{p}{n+p} log_2 \frac{p}{n+p} + \frac{n}{n+p} log_2 \frac{n}{n+p})$ $BE(A) = \sum_{i=1}^{k}\frac{p_i + n_i}{p+n} H(\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i})$ $A(BK) = H\left(\frac{p}{n+p}, \frac{n}{n+p}\right) - BE(A)$	DT, whose main working purpose is to create a tree structure of categorized data points by attribute, determines the attribute that should be at the tree root or parent level with the De Mantaras method. Where *k* denotes different values, *A* denotes the selected attribute and *E* denotes the training data, which is a subset of the training data such as $\{E_1, E_2, \dots, E_k\}$.
Naive Bayes (Pearson, 1925) (NB)	$P(A\|B) = \frac{P(B\|A)P(A)}{P(B)}$	The theoretical foundation of the strategy is Thomas Bayes' theorem, NB works according to a priori probabilities of events. Attributes are independent and the probability of each state is calculated. Classification is determined according to the highest probability value. Here, $P(A\|B)$, $P(B\|A)$, $P(A)$, and $P(B)$ denote the probability of occurrence of event A when event B occurs and the probability of occurrence of event B when event A occurs, respectively.

5.4. Success Analysis

In this study, an experimental study was conducted on three different datasets to see whether the proposed method RIG provides meaningful results. In the experiments, RIG was compared with existing feature selection approaches according to the Micro-F1 criterion with feature sizes of 50, 100, 200, 300 and 500. The best-known IG, GI, CHI2 and DFS methods were compared with RIG in terms of their impact on the performance of SVM, KNN, DT and NB classifiers. The experimental results are shown in Tables 5, Table 6, and Table 7.

Table 5 shows the Micro-F1 results for each classifier at Top-50, 100, 200, 300 and 500 dimensions on the SMS dataset. It demonstrates the classifiers' performance in several dimensions for the features chosen by the feature selection procedures. When the table is analyzed, it is seen that SVM classifier and RIG achieve the best results in all attribute sizes. Again, with KNN, it gave the best result except for dimensions 300 and 500. Similarly, the DT classifier achieved the highest values for dimensions 50, 300 and 500.

Table 5. *Micro-F1 scores results for the SMS dataset using SVM, KNN, DT, and NB*

	Accuracy results according to selected feature sizes				
SVM	50	100	200	300	500
IG	92,02	**93,54**	**91,25**	89,35	90,49
GI	93,16	91,63	**91,25**	91,25	90,49
CHI2	91,63	92,78	90,49	89,35	90,87
DFS	93,14	**93,54**	**91,25**	90,11	**91,63**
RIG	**93,25**	**93,54**	**91,25**	**91,63**	**91,63**
KNN					
IG	87,83	90,87	90,49	88,97	89,35
GI	87,45	87,83	90,78	87,07	**90,87**
CHI2	87,83	89,35	90,11	**89,35**	89,73
DFS	88,97	88,59	89,73	89,73	**90,87**
RIG	**93,93**	**92,78**	**91,25**	87,35	89,73
DT					
IG	90,87	**92,02**	**90,87**	90,11	89,35
GI	91,25	**92,02**	90,49	90,11	89,35
CHI2	91,25	**92,02**	**90,87**	90,11	89,35
DFS	91,63	90,49	**90,87**	**90,49**	89,35
RIG	**91,75**	90,97	90,35	**90,49**	**89,97**
NB					
IG	88,21	87,07	87,07	86,69	85,17
GI	84,79	88,21	85,55	84,79	85,55
CHI2	86,69	88,21	89,35	87,83	87,83
DFS	**91,63**	**93,92**	**92,78**	**93,16**	**92,78**
RIG	87,07	89,35	90,35	88,97	88,59

Table 6 shows the classifier performance according to the features presented by the feature selectors on the Eminem dataset. In the table, the RIG method, SVM, KNN and DT classifiers reached the highest value in all dimensions except dimension 100.

The performances for feature selectors on the Kat Perry dataset are shown in Table 7. The RIG method performed best with SVM in dimensions 100, 200, 300 and 500, KNN in 500, DT in all dimensions and NB in 300 and 500.

As a result, the RIG method has shown a successful performance among the existing approaches at the specified feature sizes. In particular, the RIG method performed much better with the SVM classifier.

Table 6. *Micro-F1 scores results for the Eminem dataset using SVM, KNN, DT, and NB*

	Accuracy results according to selected feature sizes				
SVM	50	100	200	300	500
IG	94,07	**94,07**	94,07	94,07	94,07
GI	94,07	**94,07**	94,07	94,07	94,07
CHI2	94,07	**94,07**	94,07	94,07	94,07
DFS	93,33	93,33	93,33	93,33	94,07
RIG	**94,81**	93,33	**94,81**	**94,81**	**95,56**
KNN					
IG	64,44	71,11	66,67	68,15	65,19
GI	68,15	68,89	68,89	68,15	69,63
CHI2	74,07	**74,81**	72,59	66,67	71,11
DFS	74,07	70,37	75,56	71,85	64,44
RIG	**77,04**	72,59	**77,78**	**76,30**	**71,85**
DT					
IG	94,07	**94,07**	94,07	94,07	94,07
GI	94,07	**94,07**	94,07	94,07	94,07
CHI2	94,07	**94,07**	94,07	94,07	94,07
DFS	93,33	93,33	93,33	93,33	93,33
RIG	**94,81**	93,33	**94,81**	**94,81**	**94,81**
NB					
IG	88,15	88,89	81,48	86,67	88,15
GI	88,15	87,41	83,70	85,19	86,67
CHI2	**93,33**	**93,33**	**93,33**	**93,33**	**94,07**
DFS	84,44	85,19	81,48	71,85	85,19
RIG	86,67	88,89	84,44	76,30	85,93

Table 7. *Micro-F1 scores results for the Kat Perry dataset using SVM, KNN, DT, and NB*

	Accuracy results according to selected feature sizes				
SVM	50	100	200	300	500
IG	**97,17**	**97,17**	**92,45**	92,45	91,51
GI	**97,17**	96,23	91,51	91,51	91,51
CHI2	**97,17**	**97,17**	**92,45**	93,40	93,40
DFS	96,23	**97,17**	**92,45**	93,40	93,40
RIG	95,28	**97,17**	**92,45**	**93,51**	**93,45**
KNN					
IG	68,87	**74,53**	68,87	**77,36**	75,47
GI	**80,19**	72,64	**76,42**	**77,36**	78,30
CHI2	69,81	**74,53**	**76,42**	76,42	75,47
DFS	70,75	69,81	75,47	76,42	76,42
RIG	72,26	71,70	68,87	70,75	**78,47**
DT					
IG	96,23	96,23	96,23	96,23	94,34
GI	96,23	96,23	96,23	96,23	96,23
CHI2	**97,17**	**97,17**	**97,17**	**97,17**	**97,17**
DFS	**97,17**	**97,17**	**97,17**	**97,17**	**97,17**
RIG	**97,17**	**97,17**	**97,17**	**97,17**	**97,17**
NB					
IG	75,28	**89,62**	92,45	83,02	77,36
GI	75,47	77,36	91,51	76,42	76,42
CHI2	**78,30**	76,42	**93,40**	75,47	73,58
DFS	68,87	71,70	**93,40**	70,75	65,09
RIG	71,32	89,43	92,51	**85,09**	**78,49**

6. CONCLUSION

A novel feature selection strategy is proposed in this paper to overcome one of the major issues in text classification, namely high dimensionality. The proposed feature selection approach utilizes the ability of Rough Sets to make effective inference in areas with incomplete and insufficient data. In addition, a hybrid structure has been created with the effective features of Rough Sets and Information Gathering. This structure distinguishes the proposed strategy from other existing filter feature selectors. Therefore, effective results were obtained in experimental studies. The proposed approach is compared with the traditional best-known method on three different textual data sets and it is observed that the approach works efficiently. Ultimately, the Rough Information Retrieval method is expected to take its place in the literature with its performance.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Aggarwal, C., & Zhai, C. (2012). A survey of text classification algorithms. In: C. C. Aggarwal, & C Zhai (Eds.), *Mining text data* (pp. 163-222). https://doi.org/10.1007/978-1-4614-3223-4_6

Alberto, T. C., Lochter, J. V., & Almeida, T. A. (2015, December 9-11). *Tubespam: Comment spam filtering on youtube*. In: Proceedings of the IEEE 14th International Conference on Machine Learning and Applications (ICMLA), Miami, Florida. https://doi.org/10.1109/ICMLA.2015.37

Bermejo, P., De la Ossa, L., G´amez, J., & Puerta, J. (2012). Fast wrapper feature subset selection in highdimensional datasets by means of filter re-ranking. *Knowledge Based Systems*, *25*(1), 35-44. https://doi.org/10.1016/j.knosys.2011.01.015

Cekik, R., & Uysal, A. K. (2020). A novel filter feature selection method using rough set for short text data. *Expert Systems with Applications*, *160*, 113691. https://doi.org/10.1016/j.eswa.2020.113691

Cekik, R., & Uysal, A. K. (2022). A new metric for feature selection on short text datasets. Concurrency and *Computation: Practice and Experience*, *34*(13), e6909. https://doi.org/10.1002/cpe.6909

Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications*, *36*(3), 5432-5435. https://doi.org/10.1016/j.eswa.2008.06.054

Chou, C., Sinha, A., & Zhao, H. (2010). A hybrid attribute selection approach for text classification. *Journal of the Association for Information Systems*, *11*(9), 491. https://doi.org/10.17705/1jais.00236

Ghareb, A., Bakar, A., & Hamdan, A. (2016). Hybrid feature selection based on enhanced genetic algorithm for text categorization. *Expert Systems with Applications*, *49*, 31-47. https://doi.org/10.1016/j.eswa.2015.12.004

Gutlein, M., Frank, E., Hall, M., & Karwath, A. (2009, March 30 - April 2). *Large-scale attribute selection using wrappers*. In: Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining, (pp. 332-339), Nashville, TN. https://doi.org/10.1109/CIDM.2009.4938668

Joachims, T. (1998, April 21-23). *Text categorization with support vector machines: Learning with many relevant features*. In: Proceedings of the European conference on machine learning (pp. 137-142). Berlin, Heidelberg. https://doi.org/10.1007/BFb0026683

Kaya, M., Bilge, H. Ş., & Yildiz, O. (2013, April 24-26). *Feature selection and dimensionality reduction on gene expressions*. In: Proceedings of the 21st Signal Processing and Communications Applications Conference (SIU) (pp. 1-4), Haspolat. https://doi.org/10.1109/siu.2013.6531476

Kaya, M., & Bilge, H. Ş. (2016, May 16-19). *A hybrid feature selection approach based on statistical and wrapper methods*. In: Proceedings of the 24th Signal Processing and Communication Application Conference (SIU) (pp. 2101-2104), Zonguldak. https://doi.org/10.1109/SIU.2016.7496186

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4), 150. https://doi.org/10.3390/info10040150

Labani, M., Moradi, P., Ahmadizar, F., & Jalili, M. (2018). A novel multivariate filter method for feature selection in text classification problems. *Engineering Applications of Artificial Intelligence*, *70*, 25-37. https://doi.org/10.1016/j.engappai.2017.12.014

Nuruzzaman, M. T., Lee, C., & Choi, D. (2011, August 31 - September 2). *Independent and Personal SMS Spam Filtering*. In: Proceedings of the IEEE 11th International Conference on Computer and Information Technology, (pp. 429-435), Paphos. https://doi.org/10.1109/CIT.2011.23

Ogura, H., Amano, H., & Kondo, M. (2009). Feature selection with a measure of deviations from Poisson in text categorization. *Expert Systems with Applications*, *36*(3), 6826-6832. https://doi.org/10.1016/j.eswa.2008.08.006

Pawlak, Z. (1998). Rough set theory and its applications to data analysis. *Cybernetics & Systems*, *29*(7), 661-688. https://doi.org/10.1080/019697298125470

Pearson, E. (1925). Bayes' theorem, examined in the light of experimental sampling. *Biometrika*, *17*(3-4), 388-442. https://doi.org/10.1093/biomet/17.3-4.388

Rehman, A., Javed, K., Babri, H. A., & Saeed, M. (2015). Relative discrimination criterion–A novel feature ranking method for text data. *Expert Systems with Applications*, *42*(7), 3670-3681. https://doi.org/10.1016/j.eswa.2014.12.013

Rehman, A., Javed, K., & Babri, H. A. (2017). Feature selection based on a normalized difference measure for text classification. *Information Processing & Management*, *53*(2), 473-489. https://doi.org/10.1016/j.ipm.2016.12.004

Rehman, A., Javed, K., Babri, H. A., & Asim, M. N. (2018). Selection of the most relevant terms based on a max-min ratio metric for text classification. *Expert Systems with Applications*, *114*, 78-96. https://doi.org/10.1016/j.eswa.2018.07.028

Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., & Wang, Z. (2007). A novel feature selection algorithm for text categorization. *Expert Systems with Applications*, *33*(1), 1-5. https://doi.org/10.1016/j.eswa.2006.04.001

Shang, C., Li, M., Feng, S., Jiang, Q., & Fan, J. (2013). Feature selection via maximizing global information gain for text classification. *Knowledge-Based Systems*, *54*, 298-309. https://doi.org/10.1016/j.knosys.2013.09.019

Sharmin, S., Shoyaib, M., Ali, A. A., Khan, M. A., & Chae, O. (2019). Simultaneous feature selection and discretization based on mutual information. *Pattern Recognition*, *91*, 162-174. https://doi.org/10.1016/j.patcog.2019.02.016

Şenol, A. (2023). Comparison of Performance of Classification Algorithms Using Standard Deviation-based Feature Selection in Cyber Attack Datasets. International Journal of Pure and Applied Sciences, *9*(1), 209-222. https://doi.org/10.29132/ijpas.1278880

Uysal, A. K., & Gunal, S. (2012). A novel probabilistic feature selection method for text classification. *Knowledge-Based Systems*, *36*, 226-235. https://doi.org/10.1016/j.knosys.2012.06.005

Wang, H., & Hong, M. (2019). Supervised Hebb rule based feature selection for text classification. *Information Processing & Management*, *56*(1), 167-191. https://doi.org/10.1016/j.ipm.2018.09.004

Wang, S., Li, D., Wei, Y., & Li, H. (2009). *A feature selection method based on fisher's discriminant ratio for text sentiment classification*. In: Proceedings of the International Conference on Web Information Systems and Mining (pp. 88-97). Berlin. https://doi.org/10.1007/978-3-642-05250-7_10

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., & Vapni, V. (2001). *Feature selection for SVMs*. Advances in neural information processing systems, Denver, CO (pp. 668-674).

Yang, Y., & Pedersen, J. O. (1997). *A comparative study on feature selection in text categorization*. 14th International Conference on Machine Learning, Nashville, USA, (pp. 412-420).

Zhang, Q., Xie, Q., & Wang, G. (2016). A survey on rough set theory and its applications. *CAAI Transactions on Intelligence Technology*, *1*(4), 323-333. https://doi.org/10.1016/j.trit.2016.11.001