Automatika

Journal for Control, Measurement, Electronics, Computing and Communications

ISSN: 0005-1144 (Print) 1848-3380 (Online) Journal homepage: https://www.tandfonline.com/loi/taut20

Two new feature selection metrics for text classification

Durmuş Özkan Şahin & Erdal Kılıç

To cite this article: Durmuş Özkan Şahin & Erdal Kılıç (2019) Two new feature selection metrics for text classification, Automatika, 60:2, 162-171, DOI: 10.1080/00051144.2019.1602293

To link to this article: https://doi.org/10.1080/00051144.2019.1602293

Published online: 15 Apr 2019.

AUTOMATIKA
2019, VOL. 60, NO. 2, 162–171
https://doi.org/10.1080/00051144.2019.1602293

Two new feature selection metrics for text classification

Durmuş Özkan Şahin and Erdal Kılıç

Department of Computer Engineering, Ondokuz Mayıs University, Samsun, Turkey

ABSTRACT

Obtaining meaningful information from data has become the main problem. Hence data mining techniques have gained importance. Text classification is one of the most commonly studied areas of data mining. The main problem about text classification is the increase in the required time and a decrease in the success of classification because of data size. To determine the right feature selection methods for text classification is the main purpose of this study. Metrics that are used frequently for feature selection like Chi-square and Information Gain were applied over different data sets and performance was measured. In this study two feature selection metrics, which are based on filtration, are recommended as alternatives to the current ones. The first recommended metric is Relevance Frequency Feature Selection metric that was obtained by adding new parameters to Relevance Frequency method that is used for term weighting in text classification. The second one is the alternative Accuracy2 metric, which was obtained by changing the parameters of Accuracy2 metric. It was observed that the suggested Relevance Frequency Feature Selection and Alternative Accuracy2 metrics offer successful results as the current metrics used frequently.

ARTICLE HISTORY
Received 24 October 2018
Accepted 26 March 2019

KEYWORDS
Text classification; text mining; feature selection; term selection

1. Introduction

The internet becomes more common as the days pass and in the meantime, smartphone and tablet use also increases. This increase in use brings an increase in the amount of data that is created and stored in text format like e-books, emails, Facebook and Twitter. There are many studies on data processing on social media [1–3]. Automatic processing of a higher volume of data and obtaining meaningful information from it will definitely make life easier. The idea of text mining, a subbranch of data mining, has emerged from this purpose and solution seeking has become important.

Text classification studies date back to 1960s but main studies have become important after 1990s and continued until today. The most important one of these studies is the expert text classification system based on rules and developed by Carnegie Group over Reuters data set [4]. The main advantage of this system is its applicability. Due to increased data size and high amount of categories, this system can analyze a limited amount of data. Because, with increasing data amount, the amount of rules also increases. As hardware components like memory and CPU become more advanced and cheaper, use of machine-learning algorithms have become more common and they were tried over text classification problems. The most important one of those algorithms is Support Vector Machine (SVM) developed by Cortes and Vapnik [5]. The most important study in which SVM is applied is the study of Joachim [6]. Rather than that one, artificial neural networks [7], Naive Bayes [8], K-Nearest Neighbors (KNN) [9] and many other algorithms are commonly used.

The main problem in text classification is the excessive size of the data. Even in a simple text classification problem, there emerge thousands of terms. Due to large features obtained from data, it takes a lot of time and memory to process this data. Because of this, it is important to choose terms that have high distinction potential rather than all terms in text classification.

1.1. Contribution and motivation

In literature, there are many feature selection metrics. Some of them are used term weighting purpose like IG and CHI, some of them are used for feature selection like Term Frequency-Inverse Document Frequency (TF-IDF) originated from the term weighting method [10–13]. Relevance Frequency (RF) is another term weighting metric. However, it is not used as a feature selection method although it gives a very accurate classification success rate in text classification problems [14]. Starting from this point, the main motivation of this study is formed by the questions of how well RF will give result when it is used as a feature selection method and whether it is possible to develop a new and

CONTACT Durmuş Özkan Şahin durmus.sahin@bil.omu.edu.tr Department of Computer Engineering, Ondokuz Mayıs University, Samsun, 55200, Turkey

alternative metric by applying some changes over some existing metrics and looking at category count instead of looking at the occurrence of terms in documents when data sets including high mount of documents.

1.2. Organization

This paper is organized as follows: In Section 2, related works are given. Existing feature selection metrics and their mathematical backgrounds are given in Section 3. In section 4, proposed methods are emphasized. In Section 5 and Section 6, used data set and experimental settings are provided respectively. Sequential Minimal Optimization (SMO) is used for classification algorithms, and the classification result is evaluated by F-score. Experimental results are discussed in Section 7. Existing method is compared to proposed systems. In Section 8, conclusions and future works are given.

2. Related works

Methods for feature selection are generally separated into three main groups [15,16] as filtered, wrapper and embedded methods. Filtered methods work independent of classification algorithm or learning model and these methods can be applied easily and in a fast way. Wrapper methods do the feature selection job over the data set with specific learning rules and intuitive way of searching. Yet, the computation cost of these methods is high and they work slowly. Due to such negative features, wrapper methods are not preferred commonly in text classification problems [17]. Rather than these two, there are also embedded methods in the literature [15,16]. These methods work in alignment with the classification algorithm during the learning phase and ensure feature selection. In terms of computation cost and time, this method has higher performance compared to wrapper methods and lower performance compared to filtered methods. Besides these three methods, the use of combined methods is also common [15].

Filtered methods are generally statistical metrics that are obtained by analyzing the number of occurrence of terms in their own categories or in opposite categories. Chi-Squared (CHI), Information Gain (IG) and Document Frequency (DF) metrics can be shown as examples of filtered methods. While choosing a feature with filtered methods, there are two different methods called locally and globally [18,19]. In some studies, local policy is called as class-based whereas global policy is called as corpus-based [10]. Local policy is a better approach for binary classification as the best keywords are found and added to the classification [10]. On the other hand, global policy is based on obtaining a unique feature vector, which is obtained via some optimization techniques, among the features that are obtained for each category. It was observed that local policy gives better results with the low amount of features whereas global policy is better with a high amount of features [18,19]. In text classification, the study made by Yang and Pedersen is a popular example of filtered feature selection [20]. It was proven that using each and every word as the feature is unnecessary and mostly, such words are not related to categories. In order to execute the process of eliminating unnecessary words, 5 metrics as DF, IG, Mutual Information (MI), CHI and Term Strength (TS) were tried over 2 different data sets. It was observed that the best results were obtained with IG and CHI metrics. On the other hand, Forman, who used local policy rule, provided a comprehensive study where he compared 20 metrics that exist at the moment with the metrics [21]. Taşçı and Güngör [18,19] conducted another comprehensive study where almost all metrics of local and global policies were compared. Additionally, in that study, 4 alternative metrics were developed for Forman's approach. Proposed methods give better results than existing methods when a less amount of feature is considered, although the overall success changes from data set to data set. Besides, some statistical distributions are also used to measure the occurrence of terms in their own categories as well as in different categories [22,23]. Ogura et al. measure the relationship between term and category by using Poisson distribution [22]. When the method using Poisson distribution is compared with IG, CHI and Gini indexes, it gives similar results with Gini index and much better results than IG and CHI metrics for the low amount of feature. Wang et al. adapted t-test, which is commonly preferred in feature selection in gene sequence studies, to text classification [23]. The proposed solution gives relatively better results for unbalanced data sets. As performance results of each metric changes and each metric have different working principles depending on the data set that is used, some combined models were created by combining different methods [24]. Zheng et al. obtained successful results by using correlation quotient and CHI together [24]. On the other hand, Neumayer et al. proposed three combined methods and tested their method over 18 different data sets [25]. The combined methods give the best results for 11 different data sets out of 18. Rehman et al. show a new feature ranking metric termed as relative discrimination criterion (RDC), which takes document frequencies for each term count of a term into account while estimating the usefulness of a term [26]. The performance of RDC is compared with four important metrics.

Wrapper methods are the ones that are based on optimization and do the intuitive search over the data set. Genetic algorithms, tabu search, and particle swarm optimization can be given as examples. These algorithms require plenty of times to be processed completely. In text classification studies, the most serious problem is a huge amount of data. When bulkiness of

intuitive algorithms combines with large vector space, an unresolvable problem emerges. Therefore, wrapper methods are not preferred directly for text classification. However, in order to benefit from the power of intuitive search, some combined methods have also been developed [27]. As wrapper methods are processed step by step in text classification, they are also called a two-step approach. Generally, the first step is a filtered method and the second one is intuitive search methods. In the first step, features are sorted with filtered methods and then some of them are eliminated based on a defined threshold value. In the second step, the intuitive search is conducted over the remaining terms and finding the best features becomes the main purpose [28]. Wang et al. suggest a bi-level feature extraction-based text mining [29]. They first apply improved Chi-Squared statistics. Then, they do a prior latent Dirichlet allocation-based feature selection [29]. Wang et al. put forward the unsupervised feature selection technique [30]. Statistical Markov model and particle swarm optimization are combined for text feature selection [30]. Other two-stage feature selection methods that are generated by combining filter-based local feature selection methods with feature transformation and wrapper-based feature selection methods were investigated in studying of Kürşat [31]. The main drawback of two-step methods is that they create an extra computation cost.

In addition to above the methods, there are many studies conducted with the expansion of deep learning. Text mining is also done with deep learning [32–35]. Although there are good results with deep learning, there are two main problems. Firstly, it is necessary to have strong hardware to design a good deep learning network. Secondly, numerous data are needed in the stage of deep learning's training. For these reasons, classical machine learning and statistical feature selection are still required in some cases.

3. Existing feature selection metrics

In this section, the mathematical background of DF, CHI, IG and Acc2 metrics that are most frequently used during feature selection in text classification will be given.

3.1. Document frequency thresholding metric

Document frequency of a term is the number of documents where that term exists. DF method is mostly preferred in text classification because the computation cost of the DF algorithm is low. DF value of each term is calculated and all terms are sorted in ascending order based on their DF values. With a pre-defined threshold value, first n terms are specified as a feature and classification step begins.

3.2. Chi-Squared metric

Chi-square test is an important nonparametric test method used to compare more than two variables for a randomly selected data. It is also known as the test of independence. It is a measure that helps to find the independence between two random variables. Chi-squared generates a value depending on the relationship between term and category, during the feature selection step in text classification. If this value is 0, then it means that there is no relationship between the term and the category. The greater this value is, the more relationship between term and category exists. The mathematical representation of chi-squared metric is shown in Equation (1).

$$\mathrm{CHI}(t_j, c_i) = N\frac{(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)} \tag{1}$$

Where a, shows the number of documents that contain the t_j term in c_i, b, shows the number of documents that do not contain the t_j term in c_i, c, shows the number of documents that contain the t_j term but do not belong to c_i, d, shows the number of documents that do not belong to c_i class and do not contain the term t_j.N, shows the total amount of documents (i.e. $N = a + b + c + d$).

CHI scores of terms are generated based on Equation (1) and sorted from high to low. First terms are selected and the classification step begins.

3.3. Information gain

This method is widely used in statistics and machine learning. Information gain uses entropy and information theory. Information gain is used for some decision tree based classifiers to work as well as it is used in feature selection by looking at the presence or absence of a term in categories. The entropy of a random X variable is expressed as:

$$H = -\sum_{x\in X} P(x)\log P(x) \tag{2}$$

where $P(x)$ represents propabiliy of an event. Information gain of the random X variable is obtained by subtracting its entropy over the whole data set from the summation of entropy values that belong to each category. The general expression of information gain is given in Equation (3).

$$IG(t) = -\sum_{i=1}^{M} P(c_i)\log P(c_i) + P(t)\sum_{i=1}^{M} P(c_i|t)\log P(c_i|t) + P(\bar{t})\sum_{i=1}^{M} P(c_i|\bar{t})\log P(c_i|\bar{t}) \tag{3}$$

where M is the number of classes, $P(c_i)$, is the probability of a document to belong to the class c_i, $P(t)$ and

$P(\bar{t})$ are the probabilities of a document in the corpus includes presence and absence of term t, $P(c_i|t)$ and $P(c_i|\bar{t})$ are the conditional probabilities of class c_i given t term presence and absence of term t, respectively.

3.4. Acc and Acc2 metrics

Acc and Acc2 are two metrics used for feature selection in local policy. Each category is shown with a different set of keywords in local policy. These metrics are developed by Forman [21]. Feature selection is done by looking at the presence distribution of a term in positive and negative categories. Besides the importance of terms with positive features, negative information is also accepted as important. With the information in negative categories, it is aimed to find out the most distinctive words between classes. In Equation (4), Acc metric is given as:

$$\mathrm{Acc}(t_j, c_i) = a - c \tag{4}$$

where a and c values are as defined in Equation (1). In this metric, if a term is frequently seen in opposite categories, c value becomes very large. Thus, the difference gets smaller. As the difference becomes very small, this term is not accepted as a good feature. When the features of any category are the things to be found out with Acc metric, the imbalance between a and c values may occur. For example, when a problem with 10 classes is considered, the a number is calculated by looking at one category and the number of c is calculated by looking at other the nine categories. Therefore, the fact that the number of large c value may affect feature selection. To solve this problem, Forman [21] proposed the Acc2 me tric. Acc2 is given as:

$$\mathrm{Acc2}(t_j, c_i) = \frac{a}{n_1} - \frac{c}{n_2} \tag{5}$$

where a and c are defined in Equation (1); n_1, and n_2 show the total number of documents in c_i class, the total number of documents that do not belong to c_i class, respectively.

4. Proposed metrics

Although there are several studies in the literature about feature selection in text classification, it is still an important matter for scientists and researchers to work on. In this part of the study, two proposed feature selection methods will be mentioned. Both methods are filter-based feature selection metrics that work within local policy.

4.1. Relevance Frequency Feature Selection (RFFS)

RF is a method that was proposed by Lan for term weighting in text classification [14]. Some metrics like IG, CHI, Odd Ratio (OR) that are used in feature selection were used as term weighting method and TF-IDF, that is originally used for term weighting, was used in feature selection [10–12]. It was observed that RF gives better results in term weighting than other methods like IG, CHI and OR [14]. Yet, the RF method, that is proposed for term weighting and gave successful results, has not been used as a feature selection method. RF is given as:

$$\mathrm{RF}(t_j, c_i) = \log\left(\frac{a}{c}\right) \tag{6}$$

Because of the fact that when the value of variable a is equal to zero the log function becomes undefined, hence Equation (6) must be turned into a new defined functions by adding constant value two. The obtained function is given in Equation (7).

$$\mathrm{RF}(t_j, c_i) = \log\left(2 + \frac{a}{c}\right) \tag{7}$$

Again, to avoid a new undefined function (i.e. the value of variable c can be zero), Equation (7) is redefined as:

$$\mathrm{RF}(t_j,\ \ c_i) = \log\left(2 + \ \frac{a}{\max(1, c)}\right) \tag{8}$$

If a term is seen frequently in its own class, the a value will be larger. On the other hand, if it is not seen at all in opposite classes of rarely seen, the c value will be smaller. According to Equation (8), as the c value gets smaller, $\frac{a}{c}$ will be larger. For terms in such cases, feature selection will be ensured by assigning high RF value. In this case, RF will find the terms that have high distinction potential between categories. For some terms, RF values may be the same. In order to make such RF values different, RF is combined with Document Frequency and Equation (9) is obtained.

$$\mathrm{RFFS}(t_j, c_i) = \mathrm{DF}(t_j, c_i) * \mathrm{RF}(t_j, c_i) \tag{9}$$

where $\mathrm{DF}(t_j, c_i)$ is the number of documents a term appears in.

4.2. Alternative Acc2 (AAcc2)

In Equation (4) and Equation (5), mathematical representations of Acc and Acc2 metrics are shared and explained before. In this section, an alternative method is proposed by making some changes in the Acc2 metric's parameters. When the data set has 10 or 20 classes, the number of documents in a negative category will be higher. In the Acc2 metric, the value of c number will increase as the number of classes in data set increases. Therefore, Forman [21] added some ratios to parameters. In the proposed method, ratios were changed as category count based, instead of document count based.

This is given in Equation (10).

$$\mathrm{AAcc2}(t_j, c_i) = a - \frac{c}{K-1} \tag{10}$$

In this equation K is the total number of categories. In Acc2, the focus is to find out how many terms exist per document; whereas in the proposed approach, the focus is term count per category. For some terms, AAcc2 values may be the same. In order to ensure da ifference in such cases, AAcc2 has been combined with Document Frequency as:

$$\mathrm{AAcc2}(t_j, c_i) = \mathrm{DF}(t_j, c_i) * \left[a - \frac{c}{K-1} \right] \tag{11}$$

5. Used data sets

In literature, the unbalanced data sets Reuters and Ohsumed, the balanced data set 20 newsgroups are mostly used. The reason why we choose this data set is to examine how metric values will work overbalanced and imbalanced data sets. Some data sets have multiple classes and multiple labels. In data sets with multiple labels, some texts may have more than one category. Inside the data sets, various contents exist from news to scientific articles.

5.1. Reuters data set

A data set, known as Reuters-21578 in the literature, that contains economics related articles. In this study, the first 10 categories that have the highest amount of documents are used. This is an example of the imbalanced data set. Because almost 40% of the data set belongs to one category. Moreover, this is a data set with multiple classes and multiple labels.

5.2. 20 newsgroups data set

A data set that is formed with approximately 20.000 documents. It includes 20 classes and it is difficult to be parsed. Some categories are very similar but some are completely different. Although some categories are very similar, their contents are not. In terms of documents existing in categories, this is a relatively balanced data set.

5.3. Ohsumed data set

Ohsumed data set was created as a subset of MEDLINE database that contains clinically based data. This data set consists of summaries of medical articles. It contains 23 categories related to article summaries about various diseases. It has multiple classes and multiple labels as well. In this study, the first 10 categories are included for classification.

6. Experimental settings

In the preprocessing step of the study, case conversion was applied over all letters in texts, non-letter characters were excluded and word parsing was done. After obtaining the words one by one, root finding step was initiated. Porter Stemmer algorithm, which was generated to find roots of English words, was used [36,37]. Words that have no importance as a feature and can be seen in any category were deleted. Lastly, TF-IDF weighting was done over words [38,39].

In the text classification application, metric values proposed in this study are also used besides successful and existing metric values like CHI and IG. In the study, during feature selection, all metrics are executed within local policy. Features from 100 to 1000 are considered and results were compared. The developed application is done in the Java programming language.

In the classification stage of the study, WEKA tool was used [40]. From classification algorithms to clustering algorithms, many different data mining methods are included in WEKA tool. In the application that was developed, SVM was preferred as it is based on linear separation principle and it is highly popular in text classification. SVM is not included in the WEKA tool but classification was made with SMO algorithm that uses SVM. In text classification studies, generally binary classification is preferred. The category to be classified is tagged as positive whereas all other categories are tagged as negative. In this study, the binary classification technique was used. There are two important reasons to use binary classification. Firstly, some data sets have multiple classes and multiple labels. In such data sets, some texts belong to multiple categories. Therefore, binary classification is the better choice. Otherwise, text that is included in multiple categories is tagged with only one category and wrong classification result is obtained. The second one is the fact that binary classification gives better results in local policy [10].

Many different methods are used in text classification studies to measure the success of classification algorithms. With the performance scale, the accuracy of belonging of a document to the related class is measured. If a sample, which is positively tagged in the data set, is classified as positive after the classification process, it is called as True Positive (TP). If a sample, which is negatively tagged in the data set, is classified as negative after the classification process, it is called True Negative (TN). If a negative sample is classified as positive, it is called False Positive (FP) and if a positive sample is classified as negative, it is called False Negative (FN). F-measure is the most frequently used performance scale in text classification. In this study, to measure the success of the classifier, F-measure was used. F-measure is the harmonic mean of precision and recall values. In Equation (12), precision value (π) and

in Equation (13), recall value (ρ) are given.

$$\pi = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{12}$$

$$\rho = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{13}$$

F-measure is given in Equation (14).

$$F = \frac{2\pi\rho}{\pi + \rho} \tag{14}$$

7. Experimental results

In this section of the study, proposed metrics and some existing metrics are tried over data sets that have different features and classification results are shared. Furthermore, words obtained from metrics are examined and similarities and differences between these words are discussed.

7.1. Comparison of features obtained via metrics

As all metrics have different working principles, features obtained via metrics are also different. Additionally, it is important for classification algorithms that which feature is selected. In Table 1, 10 best features that were obtained via different metrics and that belong to acq category in Reuters dataset are shown.

If Table 1 is analyzed, it can be seen that terms **acquir** and **acquisi** that may belong to acq category were identified by all metrics except RF. AAcc2, Acc2, and RF. could not identify the term **merger** that was identified by CHI, IG, and RFFS. Instead, AAcc2 found **march**, a term that has most probably no importance, at the top. In general, it can be stated that all metrics, except RF, generated a potential feature that is related to acq category. It was observed that features that were obtained via RF are not suitable. Eight terms out of 10 that were found by AAcc2, which was proposed as an alternative to Acc2, are already the same as Acc2's findings.

In Table 2, on the other hand, the 10 best features that are obtained by different metrics and that belong to alt-atheism category in 20 newsgroup data set are shown.

When Table 2 is examined, it can be seen that terms **atheist** and **atheism** that may belong to alt.atheism category were identified by all metrics except RF. Although it is not known whether the term **write** is important for this category or not, it is possible for it to exist in any category. This word was generated by all metrics, except CHI. The term **Islam** that may be important for this category was generated by all metrics except RF and a similar term **Christian** was found only by AAcc2 and Acc2 metrics. Besides, the verb **believe** is seen only in AAcc2 metric. Terms that were found with RF are almost completely different than other metrics' results. 8 terms out of 10 that were found by AAcc2, which was proposed as an alternative to Acc2, are already the same as Acc2's findings. RFFS metric has found 6 common terms with CHI and 7 common terms with IG, out of 10.

In Table 3, the 10 best features that were found by different metrics and that are included in C01 category in Ohsumed data set.

When Table 3 is examined, it can be seen that CHI, IG, Acc2, and AAcc2 have generated similar features whereas RF and RFFS have identified different features than the others. Because of the fact that Ohsumed dataset is a complete medical dataset, it is not clearly known which terms in Table 3 belong to which disease. 7 terms out of 10 that were found by AAcc2, which was proposed as an alternative to Acc2, are already the same as Acc2's findings. RFFS metric didn't identify common terms with either CHI or IG.

7.2. Classification successes of metrics

In Figure 1, results that were obtained from Reuters dataset with SMO algorithm are given.

When Figure 1 is examined, it can be seen that the most successful results have been obtained by CHI and IG metrics, which are already successful ones. On the other hand, RFFS gives the best result for 200, 400 and 500 features. The results that were obtained with RF are really bad. The reason is that the RF values of terms are very close to each other. With RFFS that is obtained by adding DF coefficient in front of RF, the success increased in a considerable amount. AAcc2 that was proposed as an alternative to Acc2 performed worse compared to other metrics. However, it gets closer to the other metrics as feature count is increased. When it is compared with Acc2, although it fell behind for 100 features, the closest result was obtained for an

Table 1. Features that may belong to acq category.

Metrics	CHI	IG	RF	RFFS	Acc2	AAcc2
Best Terms	acquir	compani	usair	***share***	compani	***share***
	acquisi	share	buyout	***compani***	acquir	***compani***
	stake	acquir	cyclop	***acquir***	share	***march***
	compani	offer	courier	***offer***	corp	***corp***
	merger	acquisi	undisclos	***stake***	acquisi	***offer***
	share	stake	unsolicit	***merger***	stake	***acquir***
	offer	corp	purol	***march***	offer	***stock***
	sell	merger	allegheni	***corp***	stock	***acquisi***
	common	sell	chemlawn	***acquisi***	sell	***unit***
	corp	stock	cyacq	***usair***	common	***stake***

Table 2. Features that may belong to alt.atheism category.

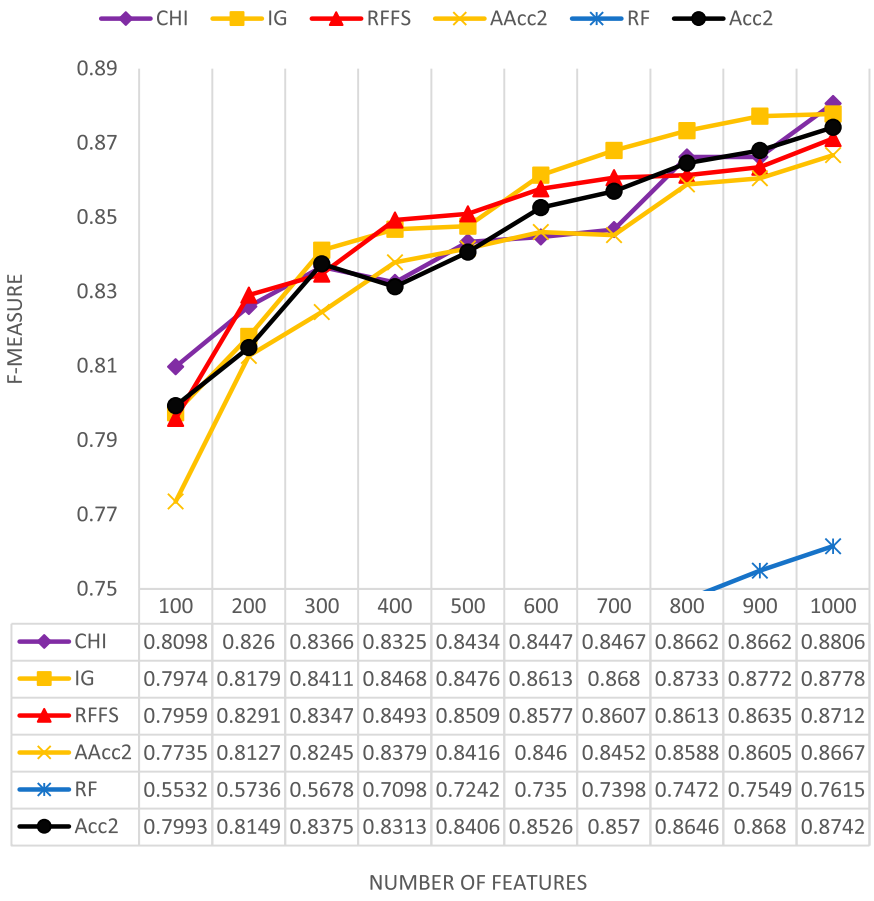
Metrics	CHI	IG	RF	RFFS	Acc2	AAcc2
Best Terms	atheist	atheist	schneider	***atheist***	atheist	***atheist***
	keith	keith	benedikt	***atheism***	write	***write***
	schneider	atheism	rushdi	***thei***	keith	***keith***
	allan	islam	mozumd	***keith***	islam	***peopl***
	atheism	moral	rosenau	***islam***	moral	***moral***
	islam	caltech	jaeger	***write***	atheism	***islam***
	livesei	livesei	buphi	***schneider***	caltech	***atheism***
	solntz	write	dbstu	***peopl***	schneider	***believ***
	caltech	schneider	beauchain	***livesei***	peopl	***christian***
	rushdi	religion	wingat	***moral***	christian	***religion***

Table 3. Features that may belong to C01 category.

Metrics	CHI	IG	RF	RFFS	Acc2	AAcc2
Best Terms	infec	infec	fungal	***gonococc***	infec	***infec***
	antibiot	antibiot	amphotericin	***purpl***	organ	***patient***
	bacteri	bacteri	antifung	***neutrocyt***	antibiot	***treatment***
	septic	organ	haemophilu	***albican***	cultur	***therapi***
	organ	cultur	nosocomi	***aminopyridin***	bacteri	***clinic***
	antimicrobi	isol	streptococc	***lancefield***	isol	***caus***
	sepsi	sepsi	mycobacterium	***immens***	caus	***cultur***
	cultur	septic	syphili	***ducreyi***	infect	***isol***
	fungal	antimicrobi	itraconazol	***coinfec***	therapi	***antibiot***
	staphylococcu	infect	bacteremia	***antepartum***	sepsi	***organ***

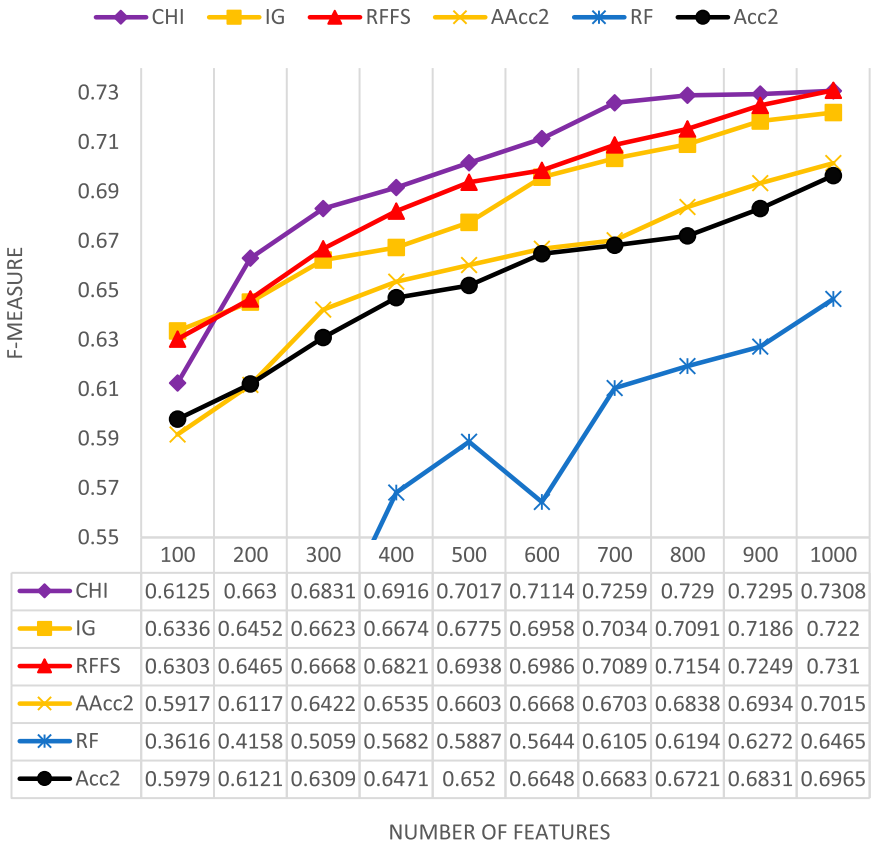
increased number of features. In Figure 2, results that have been obtained from 20 newsgroup dataset with SMO algorithm.

In Figure 2, for 100 features, RFFS gave the best result after IG. In general, the second best result was given by RFFS, after CHI. Additionally, RFFS gave the best result for 1000 features. When AAcc2 and Acc2 metrics are compared, it can be seen that AAcc2 metric gave better results in general. The reason for AAcc2 to be more successful over this dataset is the fact that it is



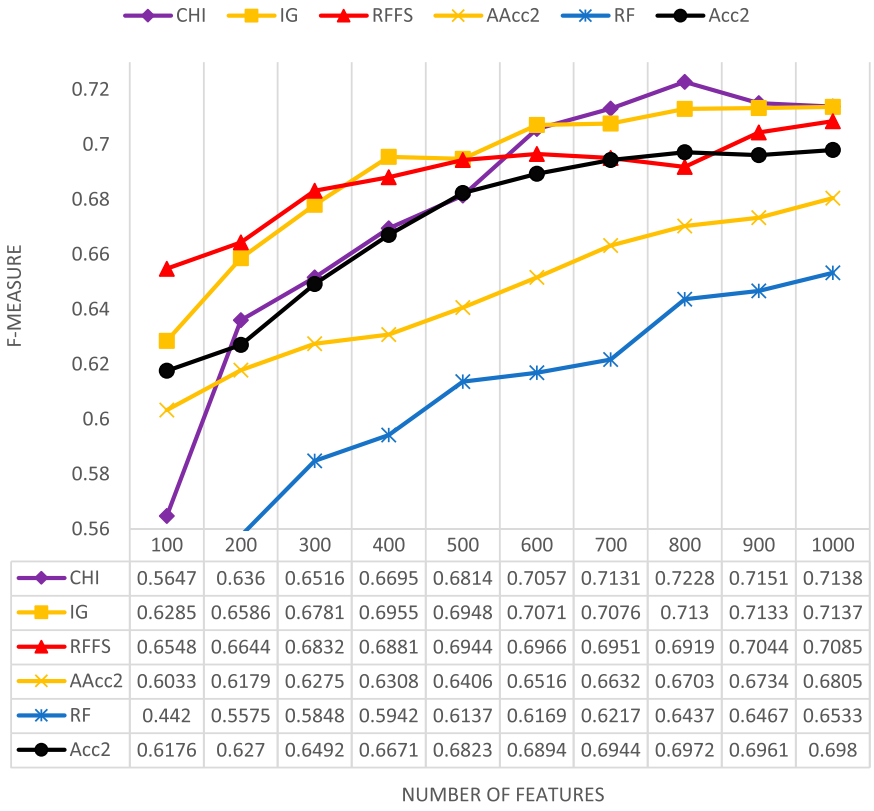
	100	200	300	400	500	600	700	800	900	1000
CHI	0.8098	0.826	0.8366	0.8325	0.8434	0.8447	0.8467	0.8662	0.8662	0.8806
IG	0.7974	0.8179	0.8411	0.8468	0.8476	0.8613	0.868	0.8733	0.8772	0.8778
RFFS	0.7959	0.8291	0.8347	0.8493	0.8509	0.8577	0.8607	0.8613	0.8635	0.8712
AAcc2	0.7735	0.8127	0.8245	0.8379	0.8416	0.846	0.8452	0.8588	0.8605	0.8667
RF	0.5532	0.5736	0.5678	0.7098	0.7242	0.735	0.7398	0.7472	0.7549	0.7615
Acc2	0.7993	0.8149	0.8375	0.8313	0.8406	0.8526	0.857	0.8646	0.868	0.8742

Figure 1. Results that were obtained from Reuters dataset with SMO algorithm.



	100	200	300	400	500	600	700	800	900	1000
CHI	0.6125	0.663	0.6831	0.6916	0.7017	0.7114	0.7259	0.729	0.7295	0.7308
IG	0.6336	0.6452	0.6623	0.6674	0.6775	0.6958	0.7034	0.7091	0.7186	0.722
RFFS	0.6303	0.6465	0.6668	0.6821	0.6938	0.6986	0.7089	0.7154	0.7249	0.731
AAcc2	0.5917	0.6117	0.6422	0.6535	0.6603	0.6668	0.6703	0.6838	0.6934	0.7015
RF	0.3616	0.4158	0.5059	0.5682	0.5887	0.5644	0.6105	0.6194	0.6272	0.6465
Acc2	0.5979	0.6121	0.6309	0.6471	0.652	0.6648	0.6683	0.6721	0.6831	0.6965

Figure 2. Results that have been obtained from 20 newsgroup dataset with SMO algorithm.



	100	200	300	400	500	600	700	800	900	1000
CHI	0.5647	0.636	0.6516	0.6695	0.6814	0.7057	0.7131	0.7228	0.7151	0.7138
IG	0.6285	0.6586	0.6781	0.6955	0.6948	0.7071	0.7076	0.713	0.7133	0.7137
RFFS	0.6548	0.6644	0.6832	0.6881	0.6944	0.6966	0.6951	0.6919	0.7044	0.7085
AAcc2	0.6033	0.6179	0.6275	0.6308	0.6406	0.6516	0.6632	0.6703	0.6734	0.6805
RF	0.442	0.5575	0.5848	0.5942	0.6137	0.6169	0.6217	0.6437	0.6467	0.6533
Acc2	0.6176	0.627	0.6492	0.6671	0.6823	0.6894	0.6944	0.6972	0.6961	0.698

Figure 3. Results that have been obtained from Ohsumed dataset with SMO algorithm.

a 20-class dataset. Because AAcc2 works based on the class count. In Figure 3, results that have been obtained from Ohsumed dataset with SMO algorithm.

When Figure 3 is analyzed, it can be seen that the most successful results for 100, 200 and 300 features were obtained with RFFS. However, as feature count

increases, there occurred no significant increase in success and it fell behind CHI and IG. In this dataset, a number of training documents are less than test documents. In such dataset, the fact that RFFS gave really good results with a low amount of features has great importance with regards to the method. When AAcc2 and Acc2 metrics are compared, it can be seen that for 100 and 200 features, similar results were obtained and Acc2 showed better performance for a higher number of features. The reason why the Acc2 metric gave better results over this dataset may be the fact that there were less training documents in this dataset than the others.

8. Conclusion and future works

In this study, two alternative and new methods were proposed for text classification. Although the methods generated different results depending on the dataset, the followings may be deducted when those results are compared with the results obtained via existing metrics that are accepted as good: RFFS gives better results than CHI when a low amount of feature is the case. Especially on Ohsumed dataset, RFFS metric shows remarkable success with a low amount of features. The proposed AAcc2 metric is an alternative to the existing Acc2 metric. An AAcc2 metric is an approach based on category count, instead of being based on document count as Acc2 metric does. In general, AAcc2 metric generates results as successful as Acc2. In the 20 Newsgroup dataset where the class count is high, AAcc2 gave a slightly better result than Acc2. For future studies, it is aimed to combine the proposed metrics and already existing metrics and to create more effective combined models. Also, the proposed metrics can be tried in some areas, such as image processing [41–43], malware analysis in the future.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] Yüksel AS, Tan FG. A real-time social network-based knowledge discovery system for decision making. Automatika. 2018;59(3–4):261–273. Available from: https://www.tandfonline.com/doi/abs/10.1080/00051144.2018.1531214

[2] Uysal AK. Feature selection for comment spam filtering on YouTube. Data Sci Appl. 2018;1(1):4–8. Available from: http://www.jdatasci.com/index.php/jdatasci/article/view/9

[3] Li L, Wu Y, Zhang Y, et al. Time+ user dual attention based sentiment prediction for multiple social network texts with time series. IEEE Access. 2019;7:17644–17653. Available from: https://ieeexplore.ieee.org/document/8628987

[4] Hayes PJ, Andersen PM, Nirenburg IB, et al. Tcs: a shell for content-based text categorization. In: Sixth Conference on Artificial Intelligence for Applications. Santa Barbara, CA: IEEE. 1990. p. 320–326. Available from: https://ieeexplore.ieee.org/document/89206

[5] Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–297. Available from: https://link.springer.com/article/10.1023/A:1022627411411

[6] Joachims T. Text categorization with support vector machines: Learning with many relevant features. In European conference on machine learning. Springer. 1998. p. 137–142. Available from: https://link.springer.com/chapter/10.1007/BFb0026683

[7] Wiener E, Pedersen JO, Weigend AS. A neural network approach to topic spotting. In Proceedings of SDAIR-95 4th annual symposium on document analysis and information retrieval. 1995. p. 317–332. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.54.6608

[8] McCallum A, Nigam K. A comparison of event models for naive Bayes text classification. In AAAI-98 workshop on learning for text categorization. 752(1):41–48; 1998. Available from: http://www.cs.cmu.edu/ knigam/papers/multinomial-aaaiws98.pdf

[9] Lam W, Ho CY. Using a generalized instance set for automatic text categorization. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. p. 81–89; 1998. Available from: https://dl.acm.org/citation.cfm?doid=290941.290961

[10] Özgür A, Özgür L, Güngör T. Text categorization with class-based and corpus-based keyword selection. In International Symposium on Computer and Information Sciences. Springer. p. 606–615; 2005. Available from: https://link.springer.com/chapter/10.1007/11569596_63

[11] Özgür A, Güngör T. Classification of skewed and homogenous document corpora with class-based and corpus-based keywords. In Annual Conference on Artificial Intelligence. Springer. p. 91–101; 2006. Available from: https://link.springer.com/chapter/10.1007/978-3-540-69912-5_8

[12] Debole F, Sebastiani F. Supervised term weighting for automated text categorization. In Text mining and its applications. Springer. p. 81–97; 2004. Available from: https://link.springer.com/chapter/10.1007/978-3-540-45219-5_7

[13] Mazyad A, Teytaud F, Fonlupt C. Information Gain Based Term Weighting Method for Multi-label Text Classification Task. In Intelligent Systems Conference; 2018. Available from: https://hal.archives-ouvertes.fr/hal-01859697

[14] Lan M, Tan CL, Su J, et al. Supervised and traditional term weighting methods for automatic text categorization. IEEE Trans Pattern Anal Mach Intell. 2009;31(4):721–735. Available from: https://ieeexplore.ieee.org/document/4509437

[15] Günal S. Hybrid feature selection for text classification. Turkish Journal of Electrical Engineering & Computer Sciences. 2012;20(2):1296–1311. Available from: https://journals.tubitak.gov.tr/elektrik/abstract.htm?id=12227

[16] Uysal AK, Gunal S. Text classification using genetic algorithm oriented latent semantic features. Expert Syst Appl. 2014;41(13):5938–5947. Available from: https://www.sciencedirect.com/science/article/pii/S0957417414001791

[17] Baccianella S, Esuli A, Sebastiani F. Using micro-documents for feature selection: The case of ordinal text classification. Expert Syst Appl. 2013;40(11):4687–4696.

Available from: https://www.sciencedirect.com/science/article/pii/S0957417413001267
[18] Taşcı S, Güngör T. An evaluation of existing and new feature selection metrics in text categorization. In Computer and Information Sciences. IEEE. p. 1–6; 2008. Available from: https://ieeexplore.ieee.org/document/4717900
[19] Taşcı Ş, Güngör T. Comparison of text feature selection policies and using an adaptive framework. Expert Syst Appl. 2013;40(12):4871–4886. Available from: https://www.sciencedirect.com/science/article/pii/S0957417413001358
[20] Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. Proceedings of the Fourteenth International Conference on Machine Learning; 1997. p. 412–420. Available from: https://dl.acm.org/citation.cfm?id = 657137
[21] Forman G. An extensive empirical study of feature selection metrics for text classification. J Mach Learn Res. 2003;3(Mar):1289–1305. Available from: http://www.jmlr.org/papers/volume3/forman03a/forman03a_full.pdf
[22] Ogura H, Amano H, Kondo M. Feature selection with a measure of deviations from Poisson in text categorization. Expert Syst Appl. 2009;36(3):6826–6832. Available from: https://www.sciencedirect.com/science/article/pii/S0957417408005484
[23] Wang D, Zhang H, Liu R, et al. . t-test feature selection approach based on term frequency for text categorization. Pattern Recognit Lett. 2014;45:1–10. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0167865514000543
[24] Zheng Z, Wu X, Srihari R. Feature selection for text categorization on imbalanced data. ACM Sigkdd Explorations Newsletter. 2004;6(1):80–89. Available from: https://dl.acm.org/citation.cfm?id = 1007741
[25] Neumayer R, Mayer R, Nørvåg K. Combination of feature selection methods for text categorization. In European Conference on Information Retrieval. Springer. 2011. p. 763–766. Available from: https://dl.acm.org/citation.cfm?id = 1996997
[26] Rehman A, Javed K, Babri HA, et al. Relative discrimination criterion–A novel feature ranking method for text data. Expert Syst Appl. 2015;42(7):3670–3681. Available from: https://www.sciencedirect.com/science/article/pii/S095741741400791X
[27] Haltas A, Alkan A, Karabulut M. Performance analysis of heuristic search algorithms in text classification. J Faculty Eng Archit Gazi Univ. 2015;30(3):417–427. Available from: http://www.mmfdergi.gazi.edu.tr/article/viewFile/5000144975/5000132374
[28] Uğuz H. A two-stage feature selection method for text categorization by using information gain, principal component analysis, and genetic algorithm. Knowl Based Syst. 2011;24(7):1024–1032. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0950705111000803
[29] Wang F, Xu T, Tang T, et al. Bilevel feature extraction-based text mining for fault diagnosis of railway systems. IEEE Trans Intell Transp Syst. 2017;18(1):49–58. Available from: https://ieeexplore.ieee.org/document/7453147/
[30] Wang Y, Wang J, Liao H, et al. Unsupervised feature selection based on Markov blanket and particle swarm optimization. J Syst Eng Electron. 2017;28(1):151–161. Available from: https://ieeexplore.ieee.org/abstract/document/7870509
[31] Uysal AK. On two-stage feature selection methods for text classification. IEEE Access. 2018;6:43233–43251. Available from: https://ieeexplore.ieee.org/abstract/document/8425702
[32] Jiang Z, Li L, Huang D, et al. Training word embeddings for deep learning in biomedical text mining tasks. IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE. p. 625–628; 2015. Available from: https://ieeexplore.ieee.org/document/7359756
[33] Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification. In Advances in neural information processing systems. p. 649–657; 2015. Available from: https://papers.nips.cc/paper/5782-character-level-convolutional-networks-for-text-classification.pdf
[34] Xu L, Jiang C, Ren Y, et al. Microblog dimensionality reduction—a deep learning approach. IEEE Trans Knowl Data Eng. 2016;28(7):1779–1789. Available from: https://ieeexplore.ieee.org/document/7430292
[35] Yao L, Mao C, Luo Y. Graph Convolutional Networks for Text Classification. arXiv preprint. 2018. arXiv:1809.05679. Available from: https://arxiv.org/abs/1809.05679
[36] Porter MF. An algorithm for suffix stripping. Program. 1980;14(3):130–137. Available from: https://dl.acm.org/citation.cfm?id = 275537.275705
[37] Porter Stemming Algorithm (PSA): http://tartarus.org/martin/PorterStemmer/ last access: October 2018.
[38] Jones KS. A statistical interpretation of term specificity and its retrieval. J Doc. 1972;28(1):11–21. Available from: https://doi.org/10.1108/eb026526
[39] Jones KS. A statistical interpretation of term specificity and its retrieval. J Doc. 2004;60(5):493–502. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi = 10.1.1.115.8343&rep = rep1&type = pdf
[40] Data Mining Software in Java (DMSJ): http://www.cs.waikato.ac.nz/ml/weka/ last access: October 2018.
[41] Khan SA, Hussain A, Usman M. Reliable facial expression recognition for multi-scale images using Weber local binary image based cosine transform features. Multimed Tools Appl. 2018;77(1):1133–1165. Available from: https://link.springer.com/article/10.1007/s11042-016-4324-z
[42] Khan SA, Ishtiaq M, Nazir M, et al. Face recognition under varying expressions and illumination using particle swarm optimization. J Comput Sci. 2018;28:94–100. Available from: https://www.sciencedirect.com/science/article/pii/S1877750317312255
[43] Khan SA, Hussain S, Xiaoming S, et al. An effective framework for driver fatigue recognition based on intelligent facial expressions analysis. IEEE Access. 2018;6:67459–67468. Available from: https://ieeexplore.ieee.org/abstract/document/8515199