Artificial Intelligence Review
https://doi.org/10.1007/s10462-021-09970-6

Feature selection methods for text classification: a systematic literature review

Julliano Trindade Pintas[1] · **Leandro A. F. Fernandes**[1] · **Ana Cristina Bicharra Garcia**[2]

Accepted: 29 January 2021

Abstract
Feature Selection (FS) methods alleviate key problems in classification procedures as they are used to improve classification accuracy, reduce data dimensionality, and remove irrelevant data. FS methods have received a great deal of attention from the text classification community. However, only a few literature surveys include them focusing on text classification, and the ones available are either a superficial analysis or present a very small set of work in the subject. For this reason, we conducted a Systematic Literature Review (SLR) that asses 1376 unique papers from journals and conferences published in the past eight years (2013–2020). After abstract screening and full-text eligibility analysis, 175 studies were included in our SLR. Our contribution is twofold. We have considered several aspects of each proposed method and mapped them into a new categorization schema. Additionally, we mapped the main characteristics of the experiments, identifying which datasets, languages, machine learning algorithms, and validation methods have been used to evaluate new and existing techniques. By following the SLR protocol, we allow the replication of our revision process and minimize the chances of bias while classifying the included studies. By mapping issues and experiment settings, our SLR helps researchers to develop and position new studies with respect to the existing literature.

Keywords Feature selection · Dimensionality reduction · Text classification · Systematic literature review

The research for this work was partially sponsored by FAPERJ (grant E-26/202.718/2018) and CNPq-Brazil (grants 311.037/2017-8 and 305.853/2018-0)

✉ Julliano Trindade Pintas
julliano@ic.uff.br

Leandro A. F. Fernandes
laffernandes@ic.uff.br

Ana Cristina Bicharra Garcia
cristina.bicharra@uniriotec.br

[1] Instituto de Computação - Universidade Federal Fluminense (UFF), Niterói, RJ, Brazil

[2] Departamento de Informática Aplicada, Universidade Federal do Estado do Rio de Janeiro (UNIRIO), Rio de Janeiro, RJ, Brazil

Published online: 24 February 2021

1 Introduction

Automated text classifiers can be used to handle several real-world problems, such as spam filtering, sentiment analysis, and news classification. Texts are usually represented by a high-dimensional and sparse document-term matrix in a space having the dimensionality of the size of the vocabulary containing word frequency counts. The high dimensionality can cause some problems, such as the curse of dimensionality and model overfitting. Feature Selection (FS) can be used to reduce dimensionality, remove irrelevant data, and increase the learning accuracy. FS is the process of automatically or manually select the features which contribute most to the classification of a given text. In text classification problems, the feature is usually some representation of a subset of words. A significant subset of features extracted from text corpora may not be relevant for the text classification task. These non-relevant features can either deteriorate the efficiency and accuracy of the classification models (Kumbhar and Mali 2013). For this reason, FS for text classification became a popular research topic in artificial intelligence and data mining conferences and journals.

Some general reviews about FS are available. Chandrashekar and Sahin (2014) and Kumar (2014) provide a general introduction to FS methods and classify them into the filter, wrapper, and embedded categories. Pereira et al. (2018) give a comprehensive survey and novel categorization of the FS techniques focusing on multi-label classification. However, these surveys did not consider in their analyses the different methods to handle the high dimensionality of the feature space, the different text representation formats such as bag of words and word embedding, and the power of the features' semantics for choosing the most efficient set of features.

FS methods have received a great deal of attention from the text classification community due to their strength in improving retrieval recall and computational efficiency (Kumbhar and Mali 2013). However important, there are only a few literature surveys (Kumbhar and Mali 2013; Shah and Patel 2016; Deng et al. 2019) that include them focusing on text classification. The ones available are either a superficial analysis or present a very small set of work in the subject. Kumbhar and Mali (2013) and Shah and Patel (2016) are more introductory studies, and both surveys don't focus only on FS methods. Besides to FS, Kumbhar and Mali (2013) address feature extraction methods and Shah and Patel (2016) address algorithms for text classification. For the best of our knowledge, there is only one review work focused exclusively on FS for text classification (Deng et al. 2019). Although Deng et al. (2019) provide a good overview of the subject, a limited proportion of published papers about FS for text classification have been included (28 studies). Among these, only fourteen were published in the last ten years, and six were published in the last five years. Besides, no clear criteria for inclusion or exclusion of the selected articles were defined. The study selection was made from other FS reviews that are not specific to text classification.

Our literature review expands existing surveys on FS methods, including up-to-date researches and providing a thorough analysis of FS methods considering the text classification task. The contribution of our literature survey lays on:

- Including a more significant number of papers covered (175 studies) resulting from a more comprehensive review in the theme;
- Bringing more up-to-date researches, including studies from 2013–2020;
- Proving a reproducible review according to an established literature review protocol;

- Providing a new research categorization for understanding the FS methods area;
- Providing a description of the experimental settings carried by the 175 reviewed studies; and
- Last but not least, we classified all 175 papers retrieved in our study according to our categorization scheme.

This paper is organized as follows: Section 2 provides background information about the main elements for text classification, including FS. The protocol of our SLR, which includes the research questions and inclusion/exclusion criteria for selecting the studies from the literature, is detailed in Sect. 3. Section 4 summarizes the issues addressed in the included studies. In Sect. 5, we cover all of the included studies by organizing them into a new categorization scheme specific to FS methods for text classification. The categorization schema proposed in this paper provides a simplified way to organize the actual methods as well as positioning new studies about FS for text categorization. The mapping of the included studies into this categorization schema allows us to identify which are the issues/topics that already have a significant amount of studies and which ones have been less explored (possibly research gaps). In Sect. 6, we survey the experiment settings used to evaluate the proposed methods. We believe that the mapping of existing studies and their experiment settings would help researchers to position and develop new studies about FS for text classification.

2 Background

Text classification is the problem to determine which class(es) a given document belongs to (Manning et al. 2008). The classification problem can be divided into three main subtypes: binary, multiclass and multilabel. If only two classes are predefined, the problem is called as a binary classification problem. If three or more classes are defined, and each document can only be associated with one of these classes, it is known as a multiclass classification problem. Finally, if each document can be simultaneously associated with two or more classes (or labels), it is defined as a multilabel classification problem.

Currently, developing models for text classification is a sophisticated process involving not only the training of models, but also numerous additional procedures, e.g., data pre-processing, transformation, and dimensionality reduction (Mirończuk and Protasiewicz 2018). This background section presents the main concepts directly related to this review's theme. Section 2.1 discusses distinct text representation models punctuating its advantages and disadvantages. Section 2.3 introduces the main concepts on FS specifically for text classification. Finally, Sect. 2.2 presents learning algorithms/architectures for text classification.

2.1 Representation models for textual data

Once you have labeled documents, the first step to construct a classification model is to extract features from text corpus. Different models of feature representation and weighting can be used for text classification and each representation model has advantages and disadvantages that must be considered. Below, we present two groups of representation models that are widely used in text classification architectures: *N*-gram based Models and Word Embedding Models.

N-gram is a set of *N* words which occurs "in that order" in a text set (Kowsari et al. 2019). The simplest and most widely used *N*-gram model is the BoW in which the $N = 1$ (called 1-gram or uni-gram model). In this model, each feature corresponds a unique word in the text. However, the *N*-gram model can also be applied with *N* values greater than 1. For example, in the 2-gram model each feature corresponds to two consecutive words. *N*-gram models with $N > 1$ could detect more information in comparison to 1-gram (Kowsari et al. 2019) because with $N = 1$ the word order information is disregarded while in 2-gram or higher models part of the word order information is captured.

In the *N*-gram model, each feature (a word or set of words) receives a value/weight for each document in the corpus. This value is usually calculated based on the frequency of that word (or set of words) in each document. The simplest is precisely the frequency of the word (or set of words) in the document, known as Term Frequency (TF). However, other weighting methods may be used. The most well-known and widely used method is the Term Frequency-Inverse Document Frequency (TF-IDF). In this method, the Inverse Document Frequency (IDF) is used in conjunction with TF in order to reduce the effect of implicitly common words in the corpus (Kowsari et al. 2019).

The *N*-gram model is usually chosen to represent text in machine learning activities due to its simplicity, robustness and the observation that simple models trained on huge amounts of data outperform complex systems trained on less data (Mikolov et al. 2013a). However, recall that *N*-gram models don't measure the semantic similarity of the words becoming a limiting factor for some types of machine learning tasks (Mikolov et al. 2013a). Thus, many researchers have been looking for representation models that capture the syntactic or semantic similarity of words (Mikolov et al. 2013a, b; Kowsari et al. 2019).

Unlike *N*-gram models that represent each word (or set of words) by a single value/weight per document, word embedding models represent each word (or set of words) by a *N*-dimension vector of real numbers (Kowsari et al. 2019). The idea behind word embedding models is that similar words have vectors with close values. In this way, the level of syntactic or semantic similarity between words can be measured based on the distance of their vectors. Different techniques for estimating word vectors have been proposed, as Word2Vec (Mikolov et al. 2013a), Glove (Pennington et al. 2014) and FastText (Bojanowski et al. 2017).

2.2 Text classification architectures

Over the years, different types of algorithms have been developed for the task of text classification (Kowsari et al. 2019). These algorithms can be divided into two main groups: traditional machine learning and deep learning. Some traditional algorithms, like Support Vector Machines (SVM), Naive Bayes (NB) and k-Nearest Neighbors (KNN), are widely studied for the text classification problem and are still commonly used by the scientific community (Kowsari et al. 2019). However, architectures based on deep learning like Convolutional Neural Network (CNN), Deep Belief Network (DBN), and Hierarchical Attention Network (HAN) are increasingly being researched for text classification (Kowsari et al. 2019). Despite having the potential to achieve excellent results in some situations, deep learning architectures have some limitations and disadvantages. Table 1 compares deep learning and traditional architecture for text classification.

Table 1 shows that each text classification architecture has advantages and disadvantages. Thus, each specific situation must be analyzed before choosing between using deep learning or traditional architecture for text classification. Two central points in

Table 1 Summary of advantages and disadvantages of text classification architectures. Adapted from Kowsari et al. (2019)

Architecture domain	Advantages	Disadvantages
Traditional	Requires a smaller amount of data	Increase the need for feature engineering
	Models are less computationally expensive to train	Other model-specific disadvantages
	Models have simpler interpretability	
Deep Learning	Reduces the need for feature engineering	Requires a large amount of data
	Can deal with complex input-output mappings	Is extremely computationally expensive to train
	Parallel processing capability	Complex model interpretability

this choice are data volume and the need to have model interpretability. Deep learning usually requires much more data than traditional machine learning algorithms and not facilitate a comprehensive theoretical understanding of learning (Kowsari et al. 2019). Therefore, if the volume of data available is small or there is a need for the interpretability of the model, the traditional architecture will probably be more suitable.

2.3 Feature selection for text classification

As shown in Sect. 2.1, the main representation models used for text classification result in high-dimensional vectors. High dimensionality can cause some problems, such as the curse of dimensionality and model overfitting. For this reason, many researchers use dimensionality reduction techniques to produce smaller feature spaces (Kowsari et al. 2019). According to Mirończuk and Protasiewicz (2018), dimensionality reduction techniques can be organized into three groups: FS, feature projection, and instance selection. While the first two types of methods aim to reduce the dimensionality of the feature space, the third aims to reduce the number of instances used for training. In this section, we focus on FS and feature projection methods.

In FS methods, the resulting feature set is a subset of the initial feature set. On the other hand, the feature projection results in a new group of features mapped from the original features. Both methods can be used in isolation or combined to reduce dimensionality. This systematic review focuses specifically on FS methods, so methods that perform feature projection are not in this study's scope.

FS methods are usually classified into three categories: filter, wrapper, and embedded (Kumar 2014). This categorization is based on the FS strategy regarding how the FS integrates into the learning activity. Filter methods are executed as a previous step and are independent of the learning activity. Wrapper methods, on the other hand, encapsulate the predictor (i.e., the classifier) and utilize the performance of the predictor to assess the relevance of features or search for the most relevant subset of features. Finally, embedded methods include FS as part of the training process.

A relevant advantage of selecting features over projecting features is because the resulting feature set is a subset of the original features. In this way, each resulting feature preserves the same meaning as the original features. This is an important point for text classification, as each feature usually represents a word or set of words. According to the survey work carried out by Mirończuk and Protasiewicz (2018), FS is the most researched dimensionality reduction technique for text classification. In our SLR, we focus specifically on FS studies for text classification.

The FS activity, which is the focus of this review, can be useful in traditional architecture and deep learning for text classification. As traditional architecture is more dependent on feature engineering activities, the selection of features has an important role in improving the classification model's accuracy. As the deep learning architecture has less dependence on feature engineering, FS tends to have less impact on the accuracy of the model. However, deep learning architectures are usually quite expensive to train. For this reason, FS may have an important utility for the deep learning architecture to reduce the computational cost.

3 Systematic literature review

The purpose of our review is to collect, organize in categories, and provide a comprehensive and recent review of FS methods for text classification. We decided to conduct a SLR to use a reproducible methodology and define explicit eligibility criteria. We aim to minimize the review bias and attempt to identify all studies that are related to our research questions.

There are several guidelines available to conduct SLRs, being Cochrane reviews protocol one of the most common in the health domain (Higgins and Green 2008). Based on Cochrane reviews protocol and other methods available in the literature, Kitchenham (2004) proposed a protocol focused on software engineering. The SLR reported in this paper follows the Kitchenham's procedures for SLR.

We have performed a SLR in three databases: (1) IEEE Xplore Digital Library, (2) ACM Digital Library, and (3) Science Direct. Our SLR protocol includes the following steps: (i) the elaboration of research questions; (ii) the definition of search strategy; (iii) high-level paper selection and classification; and (iv) detailed review of selected papers. The searches were conducted using both title and abstract. It returned a total of 1376 unique papers from journals and conference considering the past eight years (2013–2020). After abstract screening and full-text eligibility analysis, 175 studies were included in our SLR.

3.1 Research questions and search strategy

The purpose of this SLR is to find primary studies using an unbiased search strategy to answer the following research questions:

1. What are the main issues/problems that are being addressed by FS studies in text categorization task?
2. What are the different categories of methods that have been proposed?
3. What are the settings used to analyze and compare FS methods in experiments from the text categorization domain? For example: Text representation, Datasets, classifier algorithms and validation settings.

Preliminary searches were performed to assess the volume of potentially relevant studies. We identified that the query returned a small number of studies when applied only to the studies' title. Searches using full text returned an impractical volume of non-relevant studies (dozens of thousands) because the searched terms are widespread in artificial intelligence literature. Therefore we decide to perform the search using title and abstract. Additionally, in our preliminary searches, we identified the words' main variants on the concepts we are looking for. Based on that, we construct our query string:

(Feature OR Features OR Variable OR Variables OR Attribute OR Attributes) AND (Selection OR Select OR Selecting OR selected) AND (Text OR Texts OR Document OR Documents) AND (Categorization OR Classification OR Categorize OR Classify OR Categorizing OR Classifying OR Classifier)

3.2 Conducting the review

Study selection refers to the assessment of retrieved papers. For this, we defined inclusion and exclusion criteria. The first exclusion criteria specified was based on practical issues (i.e., language and date of publication). This SLR considered papers published in English and between the years 2013 and 2020. The year restriction was established considering a large number of included studies in this period. The study selection activity was executed in two steps: (1) title and abstract screening; and (2) full text screening. We performed both steps manually.

In the first screening phase, papers were included only if they contain, either in the title or in the abstract, descriptions related to *Feature Selection* and *Classification Tasks* topics in *Text Domain*.

After the first screening step, full texts were retrieved and analyzed individually. At this point, the aim was to ensure that only those studies that are related to the subject considered in this review and that are related to our research questions would be selected. The following are the main reasons for studies exclusion after the full-text analysis: (1) The study does not focus exclusively on FS (70 studies). (2) The study does not evaluate the FS method using text datasets (33 studies). (3) The study does not evaluate the classification task's method (6 studies).

The SLR reported in this paper was conducted in October 2020. Included papers reached the amount of 175 studies. Among these studies, 71 (40.57% of total) were retrieved from ACM Digital Library, 71 (40.57% of total) cames from IEEE Xplore Digital Library and 33 (18.86% of total) of them were retrieved from Science Direct. This list of papers includes journal articles and conference proceedings.

The research group that executed this SLR is composed of one D.Sc. candidate and two professors, all addressing Artificial Intelligence and Data Mining topics. Fig. 1 shows the PRISMA flow diagram for this SLR. This diagram presents a systematic review's main activities, indicating the number of studies evaluated at each stage. The PRISMA flow diagram was proposed by Moher et al. (2009) within a work that raised the preferred reporting items for systematic reviews and meta-analyses, the PRISMA statement.

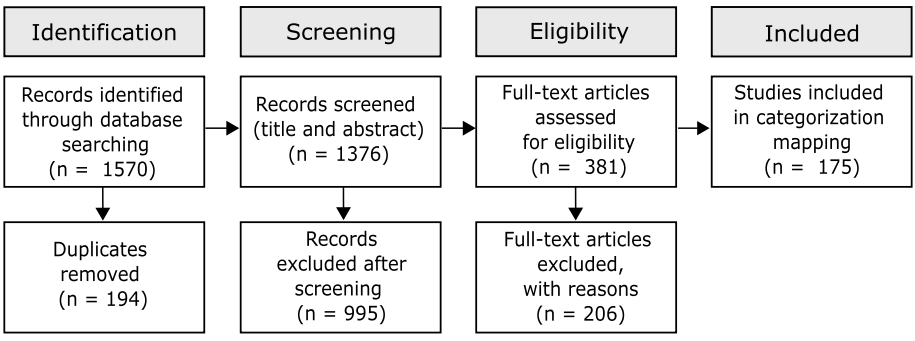

Fig. 1 PRISMA flow diagram for this systematic literature review (SLR)

4 Feature selection issues for text classification

We read and analyzed all included studies to identify the main issues that are being addressed by them (Research Question 1). After analyzing each study, we identified the main groups of problems/issues and mapped the included studies to these groups. We found that these groups of problems represent sub-tasks of the FS process (Fig. 2). They are related to each other and can be organized as:

1. **Measure feature relevance** – Measure the relevance of each feature is an essential task in FS activity. There are different ways to estimate the relevance of features, such as measuring the correlation with the target, the variable entropy, or calculating the redundancy of features (Kumar 2014). However, the basic idea is that the higher the relevance of a feature, the greater must be the power to increase the accuracy of the model (in our case, a text classifier). These studies compare existing metrics or define new metrics for calculating the predictive potential of each feature. The large part of the studies included in this review deal with issues related to the task of measure feature relevance.
2. **Subset search** – The subset search task aims to find the best subgroup of features to be used in the classification. We found two main ways to perform this search: (a) evaluating several different subsets directly in the classification activity (wrapper method), and (b) using some heuristics to assess the relevance of each subset without evaluating in a specific classifier (filter method). In both approaches, optimization methods (such as genetic algorithms or Particle Swarm Optimization (PSO)) can be used to help the search. Subset search methods commonly use as its basis some of the existing feature relevance metrics (such as Chi-square (CHI), Information Gain (IG), or Mutual Information (MI)).
3. **Globalization** – Relevance metrics and subset search methods commonly can be applied specifically for one class or label of the dataset. Therefore, a method that globalizes the results of each class/label is required to construct a final set of features that represents all classes or labels. One alternative to globalization is to use specific sets of features for

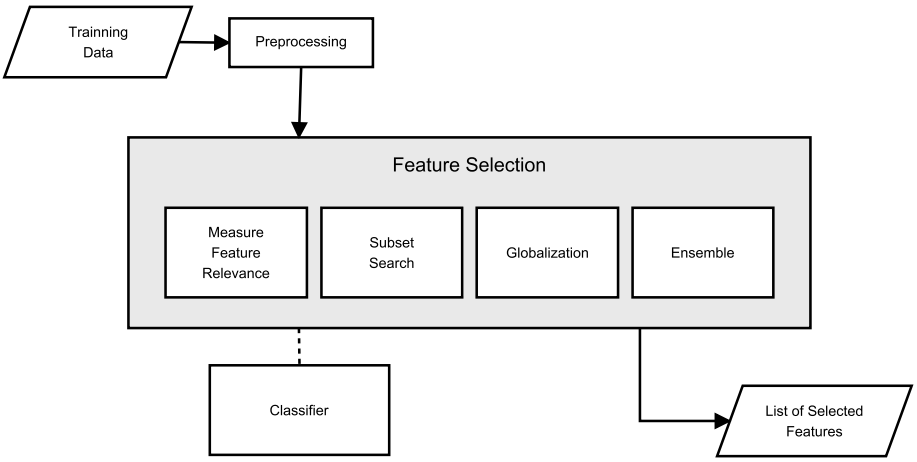

Fig. 2 The four FS sub-tasks for text classification

each class/label. However, the classifier must be designed to work this way. We mapped studies about class/label specific features in the globalization category in this review.

4. **Ensemble** – Each FS method has specific advantages and disadvantages, so combining two or more methods can lead to better results than using them separately. Ensemble studies propose or evaluate approaches to combining FS methods and/or metrics.

Sections 4.1–4.4 discuss the most relevant issues and studies for each task and present an overview of the main approaches we found to deal with each presented issue. The methods proposed in the included studies will be described in Sect. 5.

Fig. 3 presents the evolution of the number of studies by the issue group. This chart shows an increasing number of papers that address the subset search problem while a decreasing number of studies that focus on how to measure the relevance of features. Since the studies search for this review happened in October 2020, the numbers of papers for the 2020 are still partial because the source databases were still incomplete for those dates. Therefore, we mark this part of the graph in gray to show that it includes preliminary data.

4.1 Issues about measure feature relevance

After analyzing included papers that deal with the task of measure feature relevance, we mapped the most frequent issues related to this task:

- How to deal with the unbalanced or skewed datasets?
- How to avoid that rare terms receive high scores?
- How to identify and measure the redundancy of terms?
- How to consider the sparsity of the matrix?
- How to consider the position of terms?

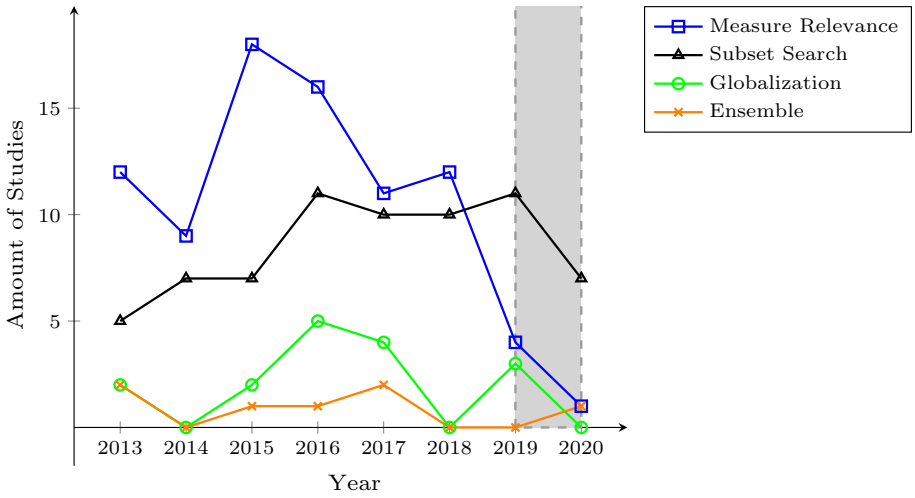

Fig. 3 Amount of studies by FS issue group over the years. The gray box indicates that data collected for 2020 may be preliminary since this review was updated in October 2020

Each one of these issues is explained next.

*Unbalanced or Skewed Datasets*For text classification, data are characterized by a large number of highly sparse terms and highly skewed categories (Rehman et al. 2018). A skewed dataset has an unbalanced class distribution, which means the number of instances in one class (majority class(es)) may be many times bigger than the number of other class instances (minority class(es)) (Japkowicz 2000). If the unbalance of classes is not treated during or before the FS, the positive features of the minority classes may receive minor relevance scores because they are present in a smaller number of documents. Consequently, the minority classes may have no features on the selected feature set. If some class/label does not have any of its features included in the final set, the classifier may not be able to classify the documents of this class/label.

*Rare Terms/Features*Several relevance metrics are based on the correlation between features and the target. In these metrics, features with high correlation receive higher scores. Because rare terms tend to be present in a few or only one class, they tend to have a higher correlation with the target. For this reason, many of the relevance metrics assign higher scores for rare features. As an example, we can cite Balanced Accuracy Measure (ACC2) and IG (Rehman et al. 2017; Wang et al. 2014a). However, rare features usually are not relevant to classification tasks, because they will have a very low likelihood of to be present in new documents.

We found two main strategies to deal with rare terms in included studies: (1) eliminating rare terms before applying the FS method; and (2) adapt the relevance metric formula to assign lower scores to rare terms.

*Redundancy Exclusion*Feature redundancy is usually defined as a high correlation among features (Wang et al. 2013). Redundant features are likely to appear simultaneously in the same documents. Considering that two features are redundant and one of them is already selected, select the second feature will contribute very little or nothing to the total relevance of the selected set. For this reason, the redundancy between features is an important factor to measure the relevance of each feature. The goal of FS is to select a highly-relevant subset with a minimum redundancy (Labani et al. 2018).

Like the case of dealing with rare features, we found two main approaches that deal with redundant features: (1) eliminating redundant features during the pre-processing phase before applying the FS method; and (2) adapt the relevance metric formula to assign lower scores to redundant features.

*Data Sparsity*For text classification, features usually consider in its recipe the frequency of words/phrases in each document. The number of possible words/phrases in all dataset training documents can be huge, and this number defines the length of the initial feature vector (Ong et al. 2015). Since each document usually includes a low percentage of the total set of words/phrases, most of the features of each document will be zero, i.e., zero frequency for words/phrases that are not present in that document.The result is a sparse composite matrix.

The matrix sparsity degrades the performance of the text classification (Ong et al. 2015). For this reason, properly handling the sparsity issue is recommended for FS methods. An alternative to address this issue is to use another representation format which does not result in a sparse matrix.

*Position/Location Inside the Text*The term location inside a text can be related to the importance of a term and, consequently, for measuring the relevance of a feature. Song et al. (2016) defines that the feature words have different capabilities to express the text in different positions of the text. Especially for news articles, information that is in the title, subtitle, or first paragraph tends to be more relevant than information on the

other paragraphs. Therefore, the location of terms within the text can be useful for FS methods.

4.2 Issues about subset search

The subset search task aims to find the best subgroup of features to be used in the classification. This search can be performed using an optimization method (such as genetic algorithms or PSO) and evaluating several different subsets directly in the classification activity (wrapper method) or using some heuristics to evaluate the relevance of each subset without evaluating in a specific classifier (filter method). Most subset search studies are focused on evaluating metaheuristics methods to improve search efficiency. Other studies focus on reducing redundancy and hyperparameter optimization.

*Search Strategy/Search Efficiency and Effectiveness*The simplest way to perform the subset search is exhaustively to evaluate all candidate subsets according to some evaluation function. However, for a dataset with N features, there exists 2^N candidate subsets. Due to a large number of features in text domain datasets, the exhaustive search usually is too costly and virtually prohibitive (Tang et al. 2014). For this reason, the main issue of the subset search for text classification is the search efficiency. Several studies included in this review proposes different metaheuristics or methods to perform the search efficiently.

Most of the included studies about subset searches are based on swarm optimization methods. In those methods, subsets of candidate features are mapped as particles, and the Swarm Optimization method tries to find the best solution (best feature subset) by exploring the search space moving the particles to find the global optimum configuration (Chopard and Tomassini 2018).

*Redundancy*Feature redundancy is usually defined in terms of some correlations within the features (Wang et al. 2013). The goal of FS is to select a highly-relevant subset with a minimum redundancy (Labani et al. 2018). Most subset search methods avoid select redundant features naturally by evaluating several combinations of features. As explained in Sect. 4.1, methods that measure the relevance of each feature in isolation need address feature redundancy explicitly.

*Feature Selection and Hyperparameter Optimization*Classification algorithms usually have several parameters whose values heavily influence its performance. Thus, determining appropriate values of parameters of a classifier is a critical issue (Ekbal and Saha 2015). Like FS, finding the best combination of parameters can be addressed as an optimization problem, called Hyperparameter Optimization. Grid search and manual search are the most widely used strategies for hyperparameter optimization (Bergstra and Bengio 2013).

Subset search methods usually are wrapper methods. That is, they use the accuracy of the classifier to evaluate each candidate subset. But notice that the predictor parameters influence the performance of the subset search. Similarly, the selected features influence hyperparameter optimization. For this reason, performing the two searches in an integrated manner may be a good approach to find the best combination of selected features and predictor parameters.

4.3 Issues about globalization

Relevance metrics and subset search methods commonly can be explicitly applied for one class or label of the dataset. Therefore, a method that globalizes the results of each class/label is required in order to construct a final set of features that represents all classes or

labels. One alternative to globalization is to use specific sets of features for each class/label. Studies about class/label specific features are mapped in the globalization category in this review.

Analyzing included studies, we found three ways to implement FS globalization within a text classification architecture:

1. Implement a local FS method for each class/label and perform the globalization subsequently.
2. Implement a global FS method designed to deal with globalization problems.
3. Adapt/Use a class/label specific classification scheme (selecting specific features for each class/label).

For the first three approaches, the main issue to be addressed is the representativeness of each class and label in the selected final set of features. For the fourth approach, the main question is how to transform the classifier or transform the problem to be able to use specific subsets of features per class/label. These globalization issues will be detailed in the following two paragraphs.

*Classes/Labels Representativeness on Final Feature Set*The basic scheme of filter-based FS assigns a score to each feature based on its discriminating power. It selects the top-k features from the feature set, where k is an empirically determined number (Agnihotri et al. 2017b). If the classification problem is multiclass or multi-label, some classes/labels may have few or no selected features in the final feature set. Therefore, a central issue for globalization is ensuring adequate representativity for all classes/labels in the final dataset.

*Class/Label Specific Features*Instead of performing the globalization and obtaining a single subset of features to be used in the classifier, it is possible to use subsets of specific features for each class or label. Some studies show this approach can improve the classification performance (Tang et al. 2016c, a). However, the classical theory as it stands requires operating in a common feature space and fails to provide any guidance for a suitable class-specific architecture (Baggenstoss 2003). Therefore, when using class/label specific features, the central issue is how to adapt the problem or the classifier to work with these class/label specific features.

In addition to the globalization approaches we identified in our systematic review, other approaches that address the globalization issue:

1. FS based on sparse learning, which focuses on the relationship between features and classes or labels (Braytee et al. 2017).
2. FS based on manifold learning (Xu et al. 2010).

4.4 Issues about ensemble

Each FS method has specific advantages and disadvantages, so combining two or more methods can lead to better results than using them separately. Ensemble studies propose or evaluate approaches by combining FS methods or metrics. We found that only seven of the included studies address the issue of ensembling FS methods. Another included studies deal with the FS methods specifically for ensembling learning approaches (for example, Boosting-based algorithms (Al-Salemi et al. 2018)), but they are not focused on ensembling FS methods. Virtually all of the included studies about the ensemble issue address

the same central problem of how to combine and aggregate the results of different FS methods. We found three main approaches to ensemble FS methods:

1. Combining selected subsets – Execute/Performs two or more FS methods isolated and then create a final set of features by combining the subsets selected by the different methods.
2. Chaining FS methods – Execute two or more FS methods in sequence, where the subset selected by a method becomes the input for the next method.
3. Ensembling rankings – Construct two or more relevance rankings (using different relevance metrics), combining the resulting rankings into a unique ranking and finally select the features using a predetermined threshold.

For each approach, different ways of combining or rankings the sets are available. The main issue of the included studies is precisely to define/find the best way to perform this combination to obtain the most relevant subset of features.

5 Feature selection methods for text classification

As explained in Sect. 4, FS for text classification should address different issues. Our SLR found that several types of methods are being proposed and evaluated to address such issues. We analyzed all the included studies of our SLR and mapped the main characteristics of each method. Based on this mapping, we designed a new categorization scheme that allows to group the methods from four different perspectives (Research Question 2): strategy, approach, target, and labeled data dependence (Fig. 4). The proposed categorization scheme helps to organize in groups and compare current methods. Additionally, it will help the positioning of future studies about FS for text classification.

The first perspective addresses the different strategies as the selection of features can be performed. It is detailed in Sect. 5.1. The different approaches (statistical, machine learning, or semantical) are mapped on the second perspective and is detailed in Sect. 5.2. The third perspective maps the type of target that the method was built to handle (binary, multiclass, multi-label, hierarchical, or ordinal). It is explained in Sect. 5.3. The fourth perspective maps the level of dependence on labeled data, being detailed in Sect. 5.4.

Each perspective is composed of a set of categories, as shown in Fig. 4. We mapped each of the studies included in this review according to each of these four perspectives of the classification schema applied on the methods described by them.Table 2 maps the issues groups described in Sect. 4 into each perspective of our categorization schema presented in this section. This table can be used to identify which strategies and approaches are being used to address each issue group. The most relevant studies on each category will be indicated during the explanation of the perspectives and respective categories in the following sections.

5.1 Categorization by strategy

As presented in Sect. 2.3, FS methods are usually classified into three categories: filter, wrapper, and embedded. The first three flows in Fig. 5 represents each one of these strategies. In this SLR, we have included two more categories to the three classical ones: Two-stages Pure and Two-stages Hybrid. Note that both strategies have the same flow in

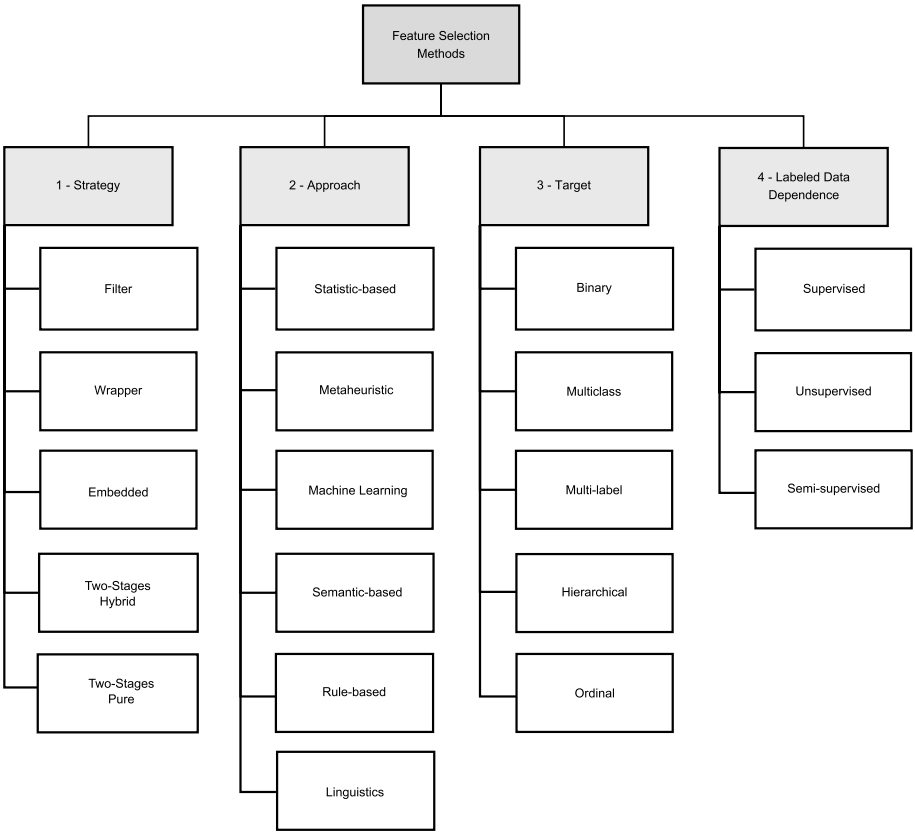

Fig. 4 Proposed categorization schema for FS methods for text classification. Each vertical represents a different categorization perspective. We used this categorization scheme to map and analyze the 175 FS methods included in this review

Fig. 5. The difference is in the choice to combine the same or different strategies. Two-stages Hybrid strategy methods combine FS methods are based on different strategies of selection. For example, the first stage may apply the filter strategy, while the second stage may use the wrapper strategy. On the other hand, some studies combine two different methods but using the same strategy. For these cases, we classify the studies in a separate category (Two-stages Pure strategy). Each one of the five strategies considered in this survey will be presented next. Fig. 6 summarizes the amount of included studies by strategy over the years.

*Filter Strategy*The main characteristic of the methods that are based on the filter strategy is to be independent of the classifier. In other words, the filter strategy does not use the performance of the classifier to assess the relevance of features or subsets of features. Lazar et al. (2012) subdivided these filter methods into two classes: ranking-based and space search. Ranking-based methods use some relevance metric to assess the predictive power of each feature, construct a ranking based on this relevance score, and apply a threshold to select the most relevant features (Chandrashekar and Sahin 2014). Space search methods aim to find the best subset of features by evaluating different combinations of features.

Table 2 Amount of FS studies for text classification by issue group and categorized according to the proposed categorization schema. All included studies were published between January/2013 and October/2020. Note: We use some abbreviations to simplify this table. TS refers to Two-Stages, and ML refers to Machine Learning

Issue group	Strategy	Approach	Target	Labeled data dependence
Measure Relevance (84 Studies)	Filter (77) TS Pure (6) Embedded (1)	Statistic-based (63) Semantic-based (9) ML (7) Linguistics (3) Rule-based (2)	Multiclass (59) Binary (19) Hierarchical (3) Ordinal (2) Multi-label (1)	Supervised (77) Semi-superv. (4) Unsupervised (3)
Subset Search (68 Studies)	Wrapper (22) Filter (21) TS Hybrid (16) Embedded (5) TS Pure (4)	Metaheuristic (30) Statistic-based (25) ML (10) Semantic-based (2) Rule-based (1)	Multiclass (43) Binary (19) Multi-label (6)	Supervised (66) Semi-superv. (1) Unsupervised (1)
Globalization (16 Studies)	Filter (14) Embedded (1) Wrapper (1)	Statistic-based (16)	Multiclass (14) Multi-label (2)	Supervised (16)
Ensemble (7 Studies)	Filter (5) TS Hybrid (1) Wrapper (1)	Statistic-based (5) Metaheuristic (1) Rule-based (1)	Binary (5) Multiclass (2)	Supervised (7)
Total (175 Studies)	Filter (117) Wrapper (24) TS Hybrid (17) TS Pure (10) Embedded (7)	Statistic-based (109) Metaheuristic (31) ML (17) Semantic-based (11) Rule-based (4) Linguistics (3)	Multiclass (118) Binary (43) Multi-label (9) Hierarchical (3) Ordinal (2)	Supervised (166) Semi-superv. (5) Unsupervised (4)

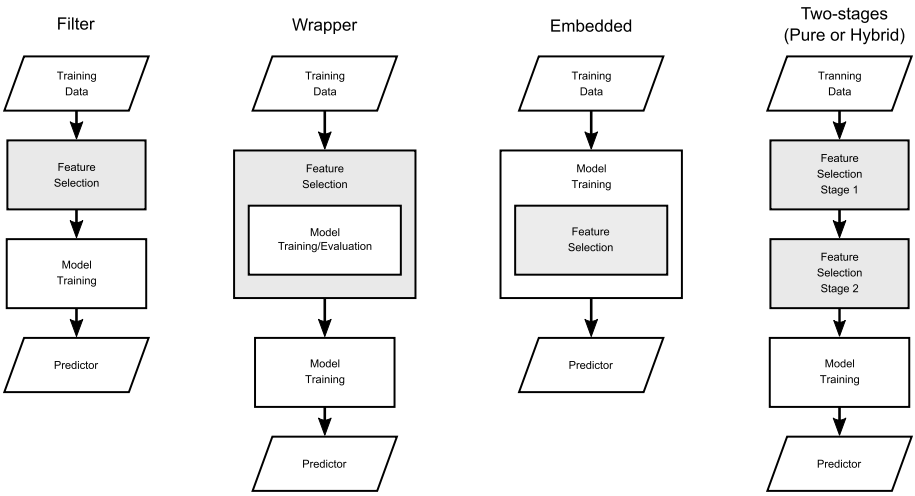

Fig. 5 Flow diagram for each Feature Selection (FS) Strategy presenting the interaction between the FS activity and the model training activity. Each of these strategies is detailed in Sect. 5.1

Table 3 summarizes studies that apply a ranking-based approach to implement the filter strategy. The studies in this table are grouped according to the base method employed to handle the problem of measuring the relevance. We found that these studies usually

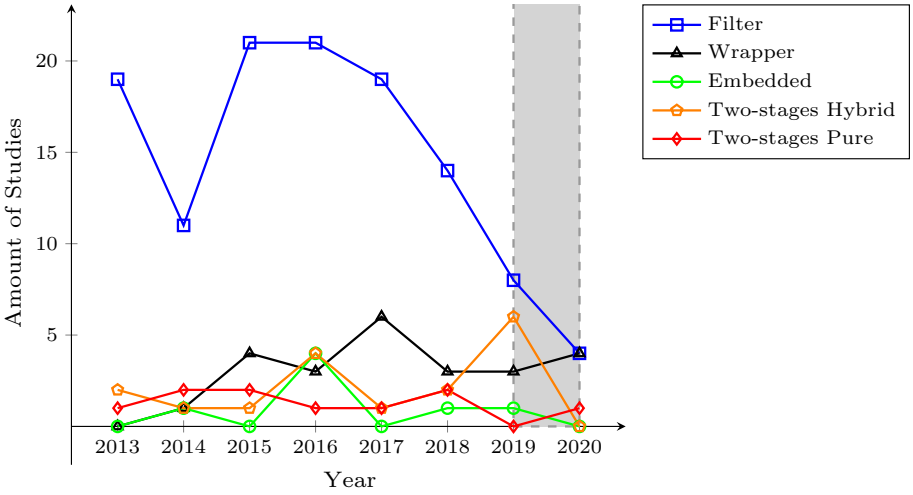

Fig. 6 Amount of FS studies by strategy over the years. The gray box indicates that data collected for 2020 may be preliminary since this review was updated in October 2020

propose an improved version of some existing relevance metrics or propose new relevance metrics. As detailed in Sect. 3, our systematic review focused on mapping studies published between 2013 and 2020. However, some studies published before this time window have proven themselves in the literature. For example, the Distinguishing Feature Selector (DFS*) method proposed by Uysal and Gunal (2012).

However, filter methods are not restricted to handle the problem of measure relevance (ranking methods). The filter strategy can also be used to handle the subset search, globalization, or ensemble problems. Table 4 presents the filter methods that address each of these issues. In Table 4, the studies are also grouped according to the base method utilized.

As described in Sect. 4.1, one of the most relevant issues for text classification is data sparsity. The FS methods generally help to reduce data sparsity by removing less relevant features. However, we found one study based on the filter strategy that mainly focuses on dealing with this issue. Ong et al. (2015) propose an improved FS metric known as Sparsity Adjusted Information Gain (SAIG), which modifies the conventional IG metric and aims to adjust the feature ranking scores according to the matrix sparsity.

Additionally, some two-stage FS methods perform the filter strategy in both stages (Two-stages Pure Strategy). Despite having two stages, these methods cannot be classified as having a hybrid strategy because they only use one strategy (the filter strategy). As an example, we can cite the study of Karabulut (2013) that presents a novel two-stage filter method based on IG theory and Geometric Particle Swarm Optimization (GPSO).

*Wrapper Strategy*Different from the filter strategy, the methods that use the wrapper strategy are dependent on the predictor because they use the performance of the predictor to evaluate the relevance of features or search for the best subset of features. For this reason, wrapper methods tend to be more computationally costly than filter methods. Most wrapper methods are based on search techniques. As explained in Sect. 4.2, the exhaustive search usually is too costly and most times prohibitive (Tang et al. 2014). For this reason, the main issue related to wrapper methods is the search efficiency (explained in

Table 3 Filter methods for measure relevance

Base method	Studies
Accuracy Measure (ACC)	Rehman et al. (2017, 2018)
Chi-square (CHI)	Fukumoto and Suzuki (2015), Agnihotri et al. (2016), Bahassine et al. (2016), Sun et al. (2017), Bahassine et al. (2018)
Class Discriminating Measure (CDM)	Ghareb et al. (2018)
Cluster-based	Yang et al. (2014), Sheydaei et al. (2015), Nam and Quoc (2016), Malji and Sakhare (2017), Chormunge and Jena (2018), Guru et al. (2020)
Comprehensively Measure Feature Selection (CMFS)	Feng et al. (2015b), Zhou et al. (2016)
Crowd-based Feature Selection (CrowdFS)	Pintas et al. (2017)
Discriminative Features Selection (DFS)	Zong et al. (2015)
Document Frequency (DF)	Chen et al. (2014), Li et al. (2014), Wang et al. (2014c), Li et al. (2015), Li (2016a, 2016b), Sarhan et al. (2016), Zhen et al. (2016), Zhou et al. (2018)
Entropy	Mladenović et al. (2016), Wu et al. (2016)
Gini Index (GI)	Chen et al. (2014), Wu et al. (2017), Ortega-Mendoza et al. (2018)
Information Gain (IG)	Chen et al. (2013), Patil and Atique (2013), Zhang et al. (2013), Gao et al. (2014), Jiang and Yu (2015), Ong et al. (2015), Wu and Xu (2016), Zhu et al. (2017), Rastogi (2018)
Latent Dirichlet Allocation (LDA)	Al-Salemi et al. (2017), Zhuang et al. (2017)
Matrix	Liang et al. (2015), Wang et al. (2016)
Mutual Information (MI)	Bagheri et al. (2013), Chen et al. (2013), Jiang and Jin (2013), Xiaoming and Tang (2013), Gunduz and Cataltepe (2015), Lifang et al. (2017)
Ontology-based	Qazi and Goudar (2018)
Part of Speech Filter (POSFilter)	Jiang and Yu (2015), Qin et al. (2016)
Postion-based	Song et al. (2016)
Relative Discrimination Criterion (RDC)	Rehman et al. (2015), Labani et al. (2018)
Rule-based	Sheydaei et al. (2015), Agnihotri et al. (2016), Ouhbi et al. (2016)
Student's *t*-Test	Pramokchon and Piamsa-Nga (2014), Wang et al. (2014a)
Term Frequency-Inverse Document Frequency (TF-IDF)	Li and Li (2015), Guru et al. (2018)
Word Embedding	Rui et al. (2016), Zhu et al. (2017), Tian et al. (2018), Lan et al. (2020)
Other Methods	Baccianella et al. (2013), Hagenau et al. (2013), Li et al. (2013a, 2013b), Ren et al. (2013), Wang et al. (2013), Baccianella et al. (2014), Badawi and Altincay (2014), Wang et al. (2014c), Yang et al. (2015), Han et al. (2016), Parlar et al. (2016), Roul et al. (2016a), Tommasel (2016), Tutkan et al. (2016), Wang et al. (2017a, 2017b), Manochandar and Punniyamoorthy (2018), Méndez et al. (2019), Islam et al. (2019), Kim and Zzang (2018), Wang and Hong (2019)

Table 4 Filter methods for subset search, globalization, and ensemble issues

Problem	Base method	Studies
Subset Search	Clustering	Song et al. (2013), Zhou et al. (2014), Roul et al. (2016b)
	Firefly Algorithm	Larabi Marie-Sainte and Alalyani (2018)
	Geometric Properties	Stambaugh et al. (2013)
	Harmony Search	Wang et al. (2014b)
	Maximum Discrimination	Tang et al. (2016b)
	Mutual Information	Tang et al. (2019), Hussain et al. (2020)
	Particle Swarm Optimization (PSO)	Karabulut (2013), Yigit and Baykan (2014)
	Rough Set	Kun and Lei (2014), Zuo et al. (2018), Cekik and Uysal (2020)
	Support Vector Machines (SVM)	Rzeniewicz and Szymanski (2013)
	Syntax Features	Vani and Gupta (2017)
	Word Embedding	Su et al. (2014), Yang and Zheng (2016), Lampos et al. (2017)
	Other Methods	Li et al. (2016a, 2016b), Tripathy et al. (2017), Guru et al. (2018)
Globalization	At Least One FeaTure (ALOFT)	Pinheiro et al. (2015), Fragoso et al. (2016)
	Chi-square (CHI)	Xu and Xu (2017)
	Class Specific Features	Tang et al. (2016c)
	Global Filter-based Feature Selection Scheme (GFSS)	Uysal (2016), Agnihotri et al. (2017b, 2018)
	Information Gain (IG)	Shang et al. (2013), Xu and Jiang (2015), Hussain et al. (2017)
	Mutual Information (MI)	Lee and Kim (2013), Agnihotri et al. (2017a)
	Other Methods	Agun and Yilmazel (2019), Yang et al. (2019)
Ensemble	Blended Feature Selection Method (BFSM)	Shen et al. (2013)
	Genetic Rank Aggregation	Onan and Korukoglu (2017)
	Hybridized term-weighting	Sabbah et al. (2016)
	Meta Feature Selection (MFS)	Li (2013)
	Square of Information Gain and Chi-square (SIGCHI)	Hai et al. (2015)

Sect. 4.2). Several studies included in this review propose different methods to improve search efficiency using metaheuristic search methods. Therefore, we identified that the wrapper strategy is commonly implemented with a metaheuristic approach. Section 5.2 details the metaheuristic approach and lists the included studies based on this approach. Wrapper methods usually are subset search methods. That is, they use the accuracy of the classifier to evaluate each candidate subset. Therefore, the predictor parameters influence the subset search performance. Similarly, the features selected influence hyperparameter optimization. For this reason, wrapper methods perform the two searches (subset search and hyperparameter search) in an integrated manner can be the best approach to find the best combination of selected features and predictor parameters. Despite the relevance of this issue, the only study included addressing it was developed by Ekbal and Saha (2015).

*Embedded Strategy*The main characteristic of embedded methods is the incorporation of the FS as part of the training process. Embedded methods aim to reduce the computation time taken up for reclassifying different subsets, which is done in wrapper methods (Chandrashekar and Sahin 2014). The embedded methods include studies that are evaluated in specific/atypical learning situations:

- Aspect-based Sentiment Analysis—Zainuddin et al. (2018).
- Class-specific Features—Tang et al. (2016c).
- Ensemble of Multi-label Classifiers—Guo et al. (2017).
- Multi-objective Genetic-Programming—Nag and Pal (2016).
- Positive and Unlabeled Learning—Zhang et al. (2014b).

Our review found that most embedded strategy studies (5 of 7) focus on the subset search issue described in Sect. 4. As embedded FS methods are part of the training algorithm, this strategy can deal efficiently with the subset search issue (Nag and Pal 2016). Additionally, we identified one embedded strategy study focused on the measure relevance issue (Naik and Rangwala 2016) and another one focused on the globalization issue by implementing class-specific features (Tang et al. 2016c). None of the embedded strategy studies in this review focus on FS ensemble issue.

Two-stages Hybrid Strategy Each of the strategies presented until now (filter, wrapper, and embedded) has specific advantages and disadvantages. For this reason, many studies explore hybrid methods that combine two different strategies into a single method. In this way, it is possible to combine their advantages and mitigate specific problems/risks. Several studies perform a filter stage before conduct the subset search to reduce the search space. For example:

- Filter Stage + Genetic Algorithm Based Search—Ghareb et al. (2016).
- Filter Stage (IG or CHI) + Rough Set—Kun and Lei (2014).
- Filter Stage + Markov Blanket Filter (MBF) Subset Search—Javed et al. (2015).
- Filter Stage + Support Vector Machine-Recursive Feature Elimination (SVM-RFE)—Zhang et al. (2014a).
- Filter Stage (CHI) + Support Vector Machine-Recursive Feature Elimination (SVM-RFE)—Chen et al. (2019).
- Filter Stage (CHI) + Particle Swarm Optimization (PSO)—Somantri et al. (2019).
- Filter Stage (MI) + Recursive Feature Elimination (RFE)—Jie and Keping (2019).
- Filter Stage (IG) + Particle Swarm Optimization (PSO)—Bai et al. (2018).
- Filter Stage (IG) + Binary Gravitational Search Algorithm (BGSA)—Kermani et al. (2019).

- Filter Stage (IG) + Improved Sine Cosine Algorithm (ISCA)—Belazzoug et al. (2020).
- Filter Stage (Ontology Filter) + Particle Swarm Optimization (PSO)—Abdollahi et al. (2019).

Two-stages Pure Strategy Some studies combine two different methods but using the same strategy. For this reason, they cannot be classified as hybrid strategy methods. Therefore, we classify them into a different strategy (Two-stages Pure). The following studies combine two stages based on filter strategy or based on wrapper strategy:

- Filter (IG or CHI or MI) + Filter (Clustering)—Ghareb et al. (2016).
- Li et al. (2013b) propose a two-step FS method. At the first step, redundancy analysis among original features based on a categorical fuzzy correlation degree is applied to filter the redundant features with a similar categorical term frequency distribution. In the second step, a conventional IG feature relevance metric is adopted to select the final feature set.
- Wrapper (Forward Feature Construction) + Wrapper (Genetic Algorithm)—Rasool et al. (2020).

5.2 Categorization by approach

During the review, we identified that FS works could be grouped according to the approach used. In this paper, the approach is related to the computational, statistical, or semantic technique used to select features. While the strategy (Sect. 5.1) defines how the method will fit into the training process and how it relates to the classifier, the approach (presented in this section) concerns the technique employed to perform the selection of features.We decided to map each method based on their primary approach since we identified that most methods do some combination of approaches, mainly with the statistic-based approach. For this reason, we did not map them into a separate category (hybrid approach).

Most published studies (109 of 175, 62.29%) use statistical metrics to measure the relevance of features and select them. These methods will be classified as statistic-based approaches. However, other studies use different approaches to select features. The main groups of approaches we have found were machine-learning-based techniques (such as clustering), semantic-based techniques, and rule-based techniques (such as Apriori). Each of these approaches will be detailed below.

Statistic-Based Approaches Gunduz and Cataltepe (2015) propose a feature relevance metric called Balanced Mutual Information (BMI) that is able to deal with the class imbalance problem through oversampling of the minority classes. They use the Synthetic Minority Oversampling Technique (SMOTE) for oversampling, which creates new minority class instances by searching for nearest neighbors of a randomly selected minority class instance. The new minority class instance value is generated by interpolation of randomly selected instances and selected neighbors of this instance. Rehman et al. (2018) propose a new feature relevance metric called Max-Min Ratio (MMR). It is a product of max-min ratios of true positives and false positives and their difference, which allows MMR to select smaller subsets of more relevant terms even in the presence of highly skewed classes.

As discussed in Sect. 4.1, there are two main strategies to deals with rare terms: (1) eliminate rare terms during the pre-processing phase before applying the FS method, and (2) adapt the relevance metric formula to assign lower scores to rare terms. Rehman et al. (2015) adopt the first strategy explicitly by removing rare features before evaluating

the proposed relevance metric called Relative Discrimination Criterion (RDC). Rehman et al. (2017) adopt the second strategy in a recent study. They propose the Normalized Difference Measure (NDM) that is an improved version of the ACC2 (Forman 2004) modified specially to assign lower relevance scores to rare features.

Labani et al. (2018) demonstrated that RDC is an effective method for identifying relevant features. A drawback is that the correlation between features is ignored, and thus RDC cannot identify redundant features. In order to mitigate this problem, Labani et al. (2018) propose the Multivariate Relative Discrimination Criterion (MRDC) that is an evolution of the RDC. Labani et al. (2018) modified the original formula to identify and measure the redundancy of features based on the correlation between then. As a result, MRDC assigns a higher relevance score to features with high discriminative power and low redundancy.

Document Frequency (DF) of a feature refers to the number of documents that include that feature. The term frequency refers to the occurrence number of a certain feature in a certain document. Most popular FS metrics for text classification such as IG, CHI, and Odds Ratio (OR), are based on DF and don't use the term frequency (Baccianella et al. 2013). However, the term frequency is a piece of important information for FS because it represents the importance of feature to each document (Wu and Xu 2016). High-Frequency terms (except stop words) that occurred in few documents are often regards as discriminators in the real-life corpus (Wang et al. 2014a).

To overcome this drawback, Baccianella et al. (2013) propose to logically break down each training document of length k into k training "micro-documents", each consisting of a single word occurrence and endowed with the same class information of the original training document. This transformation has the double effect of (a) allowing all the original FS methods based on binary information to be still straightforwardly applicable, and (b) making them sensitive to term frequency information. Wang et al. (2014a) propose a new FS metric based on term frequency and Student's t-Test. The T-TEST function is used to measure the diversity of the distributions of a term frequency between the specific category and the entire corpus. Wu and Xu (2016) propose a new FS metric that combines DF and term frequency called Limiting DF's Word Frequency. Its primary principle is summarized as follows: pre-set the threshold value of minimum DF α and the threshold value of maximum DF β, if the DF of feature word is between α and β then calculate the word frequency of this feature word or delete it otherwise.

Metaheuristic Approaches As explained in Sect. 5.1, metaheuristic search methods can be implemented to address the subset search issue and usually is combined with wrapper strategy. Metaheuristic algorithms use problem-specific heuristic information and efficiently manage the search process without exploring the whole search space (Gökalp et al. 2020). Therefore, they are ideal candidates to overcome the drawbacks of wrapper-based methods (Gökalp et al. 2020). Common meta-heuristic algorithms include the genetic algorithm and PSO (Lin et al. 2016). The included studies that implement the metaheuristic approach is listed below:

- Binary Black Hole Algorithm (BBHA)—Pashaei and Aydin (2017).
- Binary Particle Swarm Optimization (BPSO)—Shang et al. (2016).
- Cat Swarm Optimization (CSO)—Lin et al. (2016).
- Genetic Algorithm and Wrapper Approaches (GAWA)—Rasool et al. (2020).
- Improved Particle Swarm Optimization (IPSO)—Lu et al. (2015).
- Multi-Objective Automated Negotiation based Online Feature Selection (MOANOFS)—BenSaid and Alimi (2021).
- Multi-Objective Relative Discriminative Criterion (MORDC)—Labani et al. (2020).

- Memetic Feature Selection based on Label Frequency Difference (MFSLFD)—Lee et al. (2019).
- Optimized Swarm Search-based Feature Selection (OS-FS)—Fong et al. (2016).
- Small World Algorithm (SWA)—Lu and Chen (2017).
- Wrapper Feature Selection Algorithm based on Iterated Greedy (WFSAIG)—Kyaw and Limsiroratana (2019).
- Wolf Intelligence Based Optimization of Multi-Dimensional Feature Selection Approach (WI-OMFS)—Gökalp et al. (2020).

*Machine Learning-Based Approaches*Among the included studies, 17 studies use some machine learning methods directly in the FS process. These studies mainly used the following techniques:

- Clustering—Song et al. (2013), Yang et al. (2014), Zhou et al. (2014), Sheydaei et al. (2015), Nam and Quoc (2016), Roul et al. (2016b), Malji and Sakhare (2017), Chormunge and Jena (2018), Kumar and Harish (2018), Guru et al. (2020).
- SVM—Rzeniewicz and Szymanski (2013), Zhang et al. (2014a), Tripathy et al. (2017).
- Word Embedding—Yang and Zheng (2016), Lampos et al. (2017), Tian et al. (2018), Lan et al. (2020).

*Semantic-Based Approaches*Evaluate the meaning of words can be useful for FS methods because it helps to identify the relevance of words inside a text and identify the similarity between words. Among the studies included, only 11 studies use a semantic approach. Below are the semantic technologies used by each study:

- Context-capturing Features—Hagenau et al. (2013),
- Crowd-based Feature Selection (CrowdFS)—Pintas et al. (2017).
- Discriminative Personal Purity (DPP)—Ortega-Mendoza et al. (2018).
- Latent Selection Augmented Naive Bayes (LSAN)—Feng et al. (2015a).
- Ontology—Qazi and Goudar (2018), Abdollahi et al. (2019).
- Semantic Measures—Ouhbi et al. (2016).
- Semantic Similarity—Zong et al. (2015).
- Word Embedding—Su et al. (2014), Zhu et al. (2017).
- Topic Guessing—Méndez et al. (2019).

We categorize two studies (Su et al. 2014, Zhu et al. 2017) that use Word Embeddings as a semantic approach because it was used to map the meaning of the words. Both studies used Word Embedding to map the similarity of the words and perform the similarity expansion in FS. The aim of similarity expansion is to expand the set of selected features based on similarity of words.

Rule-Based Approaches Among the studies included, only four use rule-based approach. Agnihotri et al. (2016) propose a novel hybrid FS called Correlative Association Score (CAS) of terms. The CAS utilizes the concept of the Apriori algorithm to select the most informative terms. Sheydaei et al. (2015) proposed the Bit-priori Association Classification Algorithm (BACA), which combines the rule approach with a semantic approach. More recently, Wang and Hong (2019) proposes the Hebb Rule Based Feature Selection (HRFS) that assumes that terms and classes are neurons and select terms under the assumption that a term is discriminative if it keeps "exciting" the corresponding classes. Finally,

Sundararajan et al. (2020) proposes the multi-rule based ensemble FS model for sarcasm classification.

Linguistics Approaches In our review, we found three FS studies based mainly on the linguists' approach. The proposed methods use lexical or grammar information to measure the relevance of the features (Mladenović et al. 2016, Qin et al. 2016, Jiang and Yu 2015).

5.3 Categorization by target

Document classifiers may have different types of targets. Binary classifiers estimate one class for each new document within two possible categories (usually positive and negative categories). Multiclass classifiers assign each new document to one class from a list including three or more possible classes. In multi-label classification, a classifier attempts to assign multiple labels to each document, whereas a hierarchical classifier maps text onto a defined hierarchy of output categories (Mirończuk and Protasiewicz 2018). Hierarchical and ordinal classifiers can be viewed as specific types of multiclass classifiers in which classes have a relationship with each other. In the hierarchical classification, the classes are organized into hierarchical levels, whereas in ordinal classification, the classes are organized in order or sequence.

During our review, we identified that each FS method is specifically designed to work with a specific target type. The following paragraphs present the main proposed methods of each target type.

Binary Text Classification Among the 175 included studies, 43 studies (24.57%) focus on FS for binary text classification. We found that 20 studies (46.51% of 43) are related to the sentiment analysis and 6 studies (13.95% of 43) are related to spam detection. Both sentiment analysis and spam detection are usually handled as a binary classification problem. Table 5 presents the FS studies that were evaluated with binary datasets grouped by problem domain.

Multiclass Text Classification Most of studies about FS for text classification (118 of 175 included studies, 67.43%) focus on multiclass. Among these, 82 studies (69.49% of 118 studies) evaluate the method proposed using the main news classification benchmarks (datasets Reuters-21578, 20Newsgroup, Fudan, Sogou News). FS methods for multiclass or multi-label text classification need to address the globalization issue described

Table 5 FS studies using binary target text datasets

Problem domain	Studies
Sentiment Analysis	Bagheri et al. (2013), Chen et al. (2013), Ren et al. (2013), Kun and Lei (2014), Su et al. (2014), Liang et al. (2015), Ong et al. (2015), Mladenović et al. (2016), Parlar et al. (2016), Shang et al. (2016), Onan and Korukoglu (2017), Shahid et al. (2017), Tripathy et al. (2017), Yousefpour et al. (2017), Kumar and Harish (2018), Somantri et al. (2019), Gökalp et al. (2020), Islam et al. (2019)
Spam Detection	Arani and Mozaffari (2013), Wang et al. (2014c), Liu et al. (2016)v Rajamohana et al. (2017), Méndez et al. (2019)
Other Domains	Hagenau et al. (2013), Shen et al. (2013), Wang et al. (2013), Badawi and Altincay (2014), Zhang et al. (2014a, 2014b), Lin et al. (2016), Ouhbi et al. (2016), Sabbah et al. (2016), Sarhan et al. (2016), Malji and Sakhare (2017), Rehman et al. (2017), Zhuang et al. (2017); Manochandar and Punniyamoorthy (2018); Rehman et al. (2018); Abdollahi et al. (2019); Vychegzhanin et al. (2019)

in Sect. 4.3. In our review, we found studies for each implementation options described in Sect. 4.3:

1. Implement a local FS method for each class/label and subsequently perform globalization (Shang et al. 2013; Xu and Jiang 2015).
2. Implement a global FS method designed to deal with globalization problems (Lee and Kim 2013; Agnihotri et al. 2017b).
3. Adapt/Use a class/label specific classification scheme (Tang et al. 2016a, c).

*Multi-Label Text Classification*Among the included studies, only nine studies focused on FS for multi-label text classification:

- Based on supervised topic modeling for Boosting-based multi-label text categorization—Al-Salemi et al. (2017).
- Using Diversified Greedy Backward-Forward Search (DGBFS)—Ruta (2014).
- Using Ensemble Embedded Feature Selection (EEFS)—Guo et al. (2017) and Guo et al. (2019).
- Using label Pairwise Comparison Transformation (PCT) method, which converts each original multi-label sample into multiple samples with same feature vectors and different label vectors —Xu and Xu (2017).
- Using Multivariate Mutual Information (MMI)—Lee and Kim (2013).
- Using two-stage term reduction strategy based on IG theory and GPSO search—Karabulut (2013).
- Using Fuzzy Rough Feature Selection (FRFS)—Zuo et al. (2018).
- Using Memetic Feature Selection based on Label Frequency Difference (MFSLFD)—Lee et al. (2019).

As detailed in Sect. 3, our SLR protocol focused studies that explicitly state the application of text classification in the title or abstract. However, some works outside this scope are also interesting and were experimentally tested on text data. A relevant example is the mutual Information-based multi-label FS method using interaction information (Lee and Kim 2015).

*Hierarchical Text Classification*Among the included studies, only three of them focused on FS for hierarchical text classification. Naik and Rangwala (2016) investigate various filter-based FS methods for dimensionality reduction to solve the large-scale hierarchical classification problem. Lifang et al. (2017) propose a hierarchical FS method using Kullback-Leibler divergence to measure the correlation between the class and subclasses, and using MI to calculate the correlation between each feature and subclass. Song et al. (2016) propose a FS method based on category distinction and feature position information for Chinese text classification. This is the only included study in our review that deals with the issue of considering the position of words during the FS process.

Ordinal Text Classification Among the included studies, only two focused on FS for ordinal text classification. Baccianella et al. (2013) evaluate the use of micro-documents in ordinal classification. They logically break down each training document of length k into k training "micro-documents". The purpose of the use of micro-documents was explained earlier in this section. Baccianella et al. (2014) propose four novel FS metrics that have been specifically devised for ordinal classification and test them on two datasets of product review data.

5.4 Categorization by labeled data dependence

According to the Encyclopedia of Machine Learning (Sammut and Webb 2010), supervised learning refers to any machine learning process that learns a function from an input type to an output type using data comprising examples that have both input and output values. The same Encyclopedia, define unsupervised learning to any machine learning process that seeks to learn structure in the absence of either an identified output and semi-supervised learning to any machine learning process that uses both labeled and unlabeled data to perform an otherwise supervised learning or unsupervised learning task. Labeled data are data for which each object has an identified target value, the label (Sammut and Webb 2010).

Like learning methods (such as classification and regression), FS methods can also be classified into supervised, unsupervised, and semi-supervised according to their dependence on labeled data. FS methods that need labeled data can be classified as supervised method. On the other hand, FS methods that don't need labeled data can be classified as an unsupervised FS methods. Finally, FS methods that work with both labeled and unlabeled data are classified as semi-supervised.

Supervised Methods Most FS studies for text classification propose supervised methods. Considering the 175 studies included in this review, 166 (94.86% of total) are based on supervised methods. Supervised methods are mostly methods that measure the relevance of features alone or in subsets of features based on a labeled training set. Table 6 present all included studies grouped by labeled data dependence and year of publication.

Unsupervised Methods Considering the 175 studies included in this review, only four (2.29% of total) are based on unsupervised methods. In these works, three different unsupervised techniques were used:

- Term Frequency-Inverse Document Frequency (TF-IDF) and Glasgow expressions—Manochandar and Punniyamoorthy (2018) propose two modifications to the traditional TFIDF and Glasgow expressions using graphical representations to reduce the size of the feature set.
- Word Co-occurrence Matrix—Wang et al. (2016) propose an unsupervised FS algorithm through Random Projection and Gram-Schmidt Orthogonalization (RP-GSO) from the word co-occurrence matrix.
- Word Embedding—Rui et al. (2016) propose an unsupervised FS method that utilizes Word Embedding to find groups of words with similar semantic meaning. Word Embedding maps the words into vectors and remains the semantic relationships between words. After mapping the similar semantic groups, the method maintains the most representative word on behalf of the words with similar semantic meaning. Lampos et al. (2017) propose an unsupervised FS method that uses Neural Word Embeddings, trained on social media content from Twitter, to determine how strongly textual features are semantically linked to an underlying health concept.

*Semi-Supervised Methods*Semi-supervised learning uses both labeled and unlabeled data to perform an otherwise supervised learning or unsupervised learning task (Sammut and Webb 2010). Considering the 175 studies included in this review, only five (2.86% of total) studies are based on semi-supervised methods:

- Helmholtz Principle—Tutkan et al. (2016).

Table 6 FS studies grouped by labeled data dependence and year of publication

Type	Year	Studies
Supervised	2013	Arani and Mozaffari (2013), Baccianella et al. (2013), Bagheri et al. (2013), Chen et al. (2013), Hagenau et al. (2013), Imani et al. (2013), Jiang and Jin (2013), Karabulut (2013), Lee and Kim (2013), Li (2013), Li et al. (2013a, 2013b), Patil and Atique (2013), Rzeniewicz and Szymanski (2013), Shang et al. (2013), Shen et al. (2013), Song et al. (2013), Stambaugh et al. (2013), Xiaoming and Tang (2013), Zhang et al. (2013), Wang et al. (2013)
	2014	Badawi and Altincay (2014), Baccianella et al. (2014), Chen et al. (2014), Gao et al. (2014), Kun and Lei (2014), Li et al. (2014), Pramokchon and Piamsa-Nga (2014), Ruta (2014), Su et al. (2014), Wang et al. (2014c, 2014b), Yang et al. (2014), Yigit and Baykan (2014), Zhang et al. (2014a), Zhou et al. (2014)
	2015	Ekbal and Saha (2015), Feng et al. (2015a, 2015b), Fukumoto and Suzuki (2015), Gunduz and Cataltepe (2015), Hai et al. (2015), Javed et al. (2015), Jiang and Yu (2015), Li and Li (2015), Liang et al. (2015), Lu et al. (2015), Ong et al. (2015), Pinheiro et al. (2015), Rehman et al. (2015), Sheydaei et al. (2015), Xu and Jiang (2015), Yang et al. (2015), Zong et al. (2015)
	2016	Agnihotri et al. (2016), Bahassine et al. (2016), Ferreira et al. (2016), Fong et al. (2016), Fragoso et al. (2016), Ghareb et al. (2016), Li (2016a, 2016b), Li et al. (2016a, 2016b), Lin et al. (2016), Liu et al. (2016), Mladenović et al. (2016), Nag and Pal (2016), Naik and Rangwala (2016), Nam and Quoc (2016), Ouhbi et al. (2016), Parlar et al. (2016), Qin et al. (2016), Roul et al. (2016a, 2016b), Sabbah et al. (2016), Sarhan et al. (2016), Shang et al. (2016), Song et al. (2016), Tang et al. (2016b, 2016c), Tang and He (2016), Tommasel (2016), Uysal (2016), Wu et al. (2016), Wu and Xu (2016), Yang and Zheng (2016), Zhen et al. (2016), Zhou et al. (2016)
	2017	Agnihotri et al. (2017a, 2017b), Al-Salemi et al. (2017), Chen et al. (2017), Fragoso et al. (2017), Guo et al. (2017), Hussain et al. (2017), Kumbhar et al. (2017), Lifang et al. (2017), Lu and Chen (2017), Malji and Sakhare (2017), Onan and Korukoglu (2017), Pintas et al. (2017), Rajamohana et al. (2017), Rehman et al. (2017), Shahid et al. (2017), Sun et al. (2017), Tripathy et al. (2017), Vani and Gupta (2017), Wang et al. (2017a), Wu et al. (2017), Xu and Xu (2017), Yousefpour et al. (2017), Zhu et al. (2017), Zhuang et al. (2017)
	2018	Bai et al. (2018), Benitez et al. (2018), Canuto et al. (2018), Chen et al. (2018), Chormunge and Jena (2018), Ghareb et al. (2018), Guru et al. (2018, 2018), Kumar and Harish (2018), Labani et al. (2018), Larabi Marie-Sainte and Alalyani (2018), Nogueira Rios and Gama Bispo (2018), Ortega-Mendoza et al. (2018), Qazi and Goudar (2018), Rastogi (2018), Rehman et al. (2018), Bahassine et al. (2018), Tian et al. (2018), Zainuddin et al. (2018), Zhou et al. (2018), Zuo et al. (2018)

Table 6 (continued)

Type	Year	Studies
	2019	Abdollahi et al. (2019), Agnihotri et al. (2018), Agun and Yilmazel (2019), Belazzoug et al. (2020), Chen et al. (2019), Guo et al. (2019), Islam et al. (2019), Jie and Keping (2019), Kermani et al. (2019), Kim and Zzang (2018), Kyaw and Limsiroratana (2019), Lee et al. (2019), Méndez et al. (2019), Somantri et al. (2019), Tang et al. (2019), Vychegzhanin et al. (2019), Wang and Hong (2019), Yang et al. (2019)
	2020	Gökalp et al. (2020), Labani et al. (2020), Lan et al. (2020), BenSaid and Alimi (2021), Cekik and Uysal (2020), Guru et al. (2020), Hussain et al. (2020), Rasool et al. (2020), Sundararajan et al. (2020)
Unsupervised	2016	Rui et al. (2016), Wang et al. (2016)
	2017	Lampos et al. (2017)
	2018	Manochandar and Punniyamoorthy (2018)
Semi-Supervised	2013	Ren et al. (2013)
	2014	Zhang et al. (2014b)
	2016	Han et al. (2016), Tutkan et al. (2016)
	2017	Wang et al. (2017b)

– Information Theory—Wang et al. (2017b).
– Positive and Unlabeled Learning—Zhang et al. (2014b), Han et al. (2016).
– Pseudo Labels—Ren et al. (2013).

6 Experiment settings analysis

The studies included in this review use different combinations of experiment settings, such as different datasets, classification algorithms, and performance metrics. Due to a large number of studies and the consequently large amount of experiment's settings used, define the ideal setting for a new experiment can be very challenging.

The aim of this section is mapping and summarizing the settings of the experiments that are being used to analyze and compare FS methods for text categorization (Research Question 3). We focus on analyzing the following settings:

– What text representation are being used? (Sect. 6.1)
– What public datasets, language of text corpora in datasets, and dataset domains are being used to evaluate the methods? (Sect. 6.2)
– What classifier algorithms are being used to evaluate the effectiveness of FS methods? (Section 6.3)
– Which validation settings are the most used? (Sect. 6.4)

We aim to help the design of new researches by providing a summary of which experiment settings are being used. Additionally, we have identified which settings are desirable and are underutilized.

6.1 Text representation used in experiments

Textual data can be represented in different formats for text classification. In Sect. 2, we present the widely used *N*-gram and Word Embedding representation models. Considering the works included in this review, Table 7 shows that 88.57% of the methods were evaluated using exclusively Bag of Words (BoW) (uni-gram). Among the remaining works, no other mode of representation was found to be prevalent. It is interesting to note that eight studies used the combination of different representations, two of which combining BoW and Word Embedding.

6.2 Datasets used in experiments

The primary way to evaluate the effectiveness/efficiency of a FS method is training and measuring the performance of a classifier using the FS method. In this section, we indicate which public datasets are most commonly used in FS studies for text classification. We also map the most frequently used languages and domains.

Public Datasets Most of the papers included in this review used public datasets to evaluate the proposed methods. Few studies have used private or specifically collected datasets. The use of public datasets is recommended because it facilitates the comparison of methods. Table 8 presents the most used public datasets. As our review mapped a considerable number of studies and each one can use several different datasets, a list of all datasets would be very long and would mostly include datasets that were used by

Table 7 Text representation models to evaluate FS methods

Representation	Number of studies	Example references
Bag of Words (BoW) (Uni-gram)	155	Guru et al. (2018) Uysal (2016) Rehman et al. (2018)
BoW (Uni-gram) + Part of Speech (POS)	2	Jiang and Yu (2015) Rasool et al. (2020)
BoW (Uni-gram) + Termset	1	Badawi and Altincay (2014)
BoW (Uni-gram) + Word Embeddings	2	Lampos et al. (2017) Zhu et al. (2017)
N-gram ($N > 1$)	1	Agnihotri et al. (2016)
N-gram + Part of Speech (POS)	1	Zainuddin et al. (2018)
POS + Chunk based Features	1	Vani and Gupta (2017)
POS + Lexicon + Word Embedings	1	Su et al. (2014)
POS-Pattern (3-gram)	1	Yousefpour et al. (2017)
Word Embeddings	2	Tian et al. (2018) Lan et al. (2020)
Bag of Discriminative Words (BoDW)	1	Zhuang et al. (2017)
Dense word co-occurrence matrix	1	Wang et al. (2016)
Meta-features	1	Canuto et al. (2018)
Context specific features	5	Hagenau et al. (2013) Tommasel (2016) Li et al. (2016a) Ekbal and Saha (2015) Sundararajan et al. (2020)
Total	175	–

a single study. For this reason, we focus on mapping and presenting in Table 8 only the datasets that are public and that were used by at least two studies mapped in our review.

Language of Text Corpora in Datasets The majority of the papers included in this review (72.57%) used only English text corpora to evaluated their FS methods. The second most used language is Chinese (26 studies). The third language is Arabic (seven studies). Only four studies perform experiments using two different languages (English and Chinese). Wang and Hong (2019) were the only ones who used three different datasets languages (English, Turkish and Kurdish Sorani) in the same study. Table 9 presents the languages with at least two studies utilizing that language in their datasets.

The following languages were considered by only one study: Serbian (Mladenović et al. 2016), Hinglish (Ravi and Ravi 2016), Indian (Trivedi and Tripathi 2017), Tibetan (Jiang and Yu 2015), Vietnamese (Hai et al. 2015), Japanese (Fukumoto and Suzuki 2015), Russian (Vychegzhanin et al. 2019), Idonesian (Somantri et al. 2019), Italian (Ferilli et al. 2015), and Malay (Alshalabi et al. 2013).

Table 8 Most commonly used public datasets to evaluate FS methods

Dataset	Domain	Language	Studies	Percentage (%)
Reuters-21578	News	English	57	32.57
20NewsGroup	News	English	40	22.86
WebKB	Web Content	English	19	10.86
Oshumed	Medical	English	12	6.86
Fudan	News/Web Content	Chinese	8	4.57
TDT/TDT2	News	English	8	4.57
TREC	Open Domain Questions	English	8	4.57
WAP	Web Content	English	7	4.00
Sogou	News	Chinese	6	3.43
Sector	Web Content	English	4	2.29
UCI Datasets	Several domains	English	4	2.29
Enron	Email (Spam)	English	3	1.71
k1a/k1b	Web Content	English	3	1.71
RCV1	News	English	2	1.14

Table 9 Most used language of text corpora in datasets to evaluate FS methods

Language	Studies	Percentage (%)
English	127	72.57
Chinese	26	14.86
Arabic	7	4.00
Persian	2	1.14
Turkish	2	1.14
English and Chinese	4	2.29

6.3 Classification algorithms used in experiments

The studies included in this review propose new or improved FS methods for text classification. To evaluate the performance of the proposed method, the authors perform the classification task using one or more classification algorithms. The choice of the classification algorithms for the experiment directly impacts the classification result and, therefore, the evaluation of the proposed method. Table 10 and Fig. 7 present the most used classification algorithms in studies experiments.

The most used algorithms are NB and SVM because they are recognized as having good results in the task of classifying texts (Agnihotri et al. 2017b). Table 11 presents the distribution of studies by the number of algorithms used.

6.4 Validation settings used in experiments

When designing a new experiment, the scientists must clearly define the validation method and whether any statistical tests will be performed to refute or not their hypotheses. This section presents the main evaluation settings used in studies included in this review.

Table 10 Classifiers that are most often used to evaluate FS methods

Algorithm	Studies	Percentage (%)
Support Vector Machines (SVM)	103	58.86
Naive Bayes (NB)	99	56.57
k-Nearest Neighbors (KNN)	45	25.71
Decision Tree (DT)	22	12.57
Random Forest (RF)	11	6.29

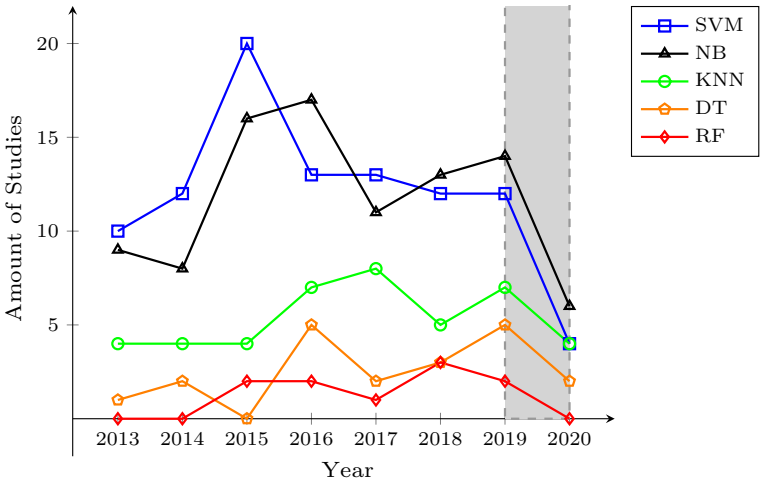

Fig. 7 Classifiers that have been most often used to evaluate FS methods over the years. The gray box indicates that data collected for 2020 may be preliminary since this review was updated in October 2020

Table 11 Number of classifiers used to evaluate FS methods

Number of tested classifiers	Studies	Percentage (%)
1 classifier	89	50.86
2 classifiers	43	24.57
3 classifiers	21	12.00
4 classifiers	15	8.57
5 or more classifiers	7	4.00
Total	175	100

Validation Method To evaluate the proposed method, classification algorithms need to be trained and tested using different datasets. This is usually done by:

– Performing *k*-fold cross-validation. In cross-validation, the data is partitioned into *k* subsets, called folds. The learning algorithm is then applied *k* times, each time one different fold is selected as the test set, and the remaining are used as the training set (Sammut and Webb 2010).

Table 12 Validation methods used in experiments

Validation method	Studies	Percentage (%)
10-Fold Cross-validation	55	31.43
5-Fold Cross-validation	15	8.57
4-Fold Cross-validation	2	1.14
3-Fold Cross-validation	4	2.29
Random Cross-validation	1	0.57
40% Train + 60% Test	1	0.57
50% Train + 50% Test	13	7.43
60% Train + 40% Test	2	1.14
65% Train + 35% Test	1	0.57
67% Train + 33% Test	2	1.14
70% Train + 30% Test	12	6.86
75% Train + 25% Test	3	1.71
80% Train + 20% Test	3	1.71
90% Train + 10% Test	2	1.14
Dataset Original Split	33	18.86
Time Split	2	1.14
Variable Length Training Set	2	1.14
Not Described	22	12.57
Total	175	100

– Splitting the dataset into two different sets (training and test sets). Some studies use the standard split between training and testing available in some public datasets. Other studies define their criteria for this division. The most common is the division based on predefined percentages. However, some studies perform division based on other criteria, such as time division or varying the size of each set within a predefined range.

Approximately half (44.00%) of the studies covered in this review were cross-validated and the other half (43.43%) used different sets of training and testing. Table 12 and Fig. 8 present the validation methods used.

Statistical Significance Test The machine learning community has become increasingly aware of the need for statistical validation of the published results (Demšar 2006). Studies covered in this survey usually evaluate the efficacy of the proposed methods by comparing the proposed solution to other FS methods. The purpose of the comparison is to verify whether the use of the proposed method increases the accuracy/precision/coverage of the classification activity in contrast to the other FS methods. Although practically all studies performed comparisons to demonstrate an improvement in classification performance, we identified that only 29.71% of them used some statistical method to confirm the statistical significance of the results. Table 13 shows which statistical methods have been used for this purpose.

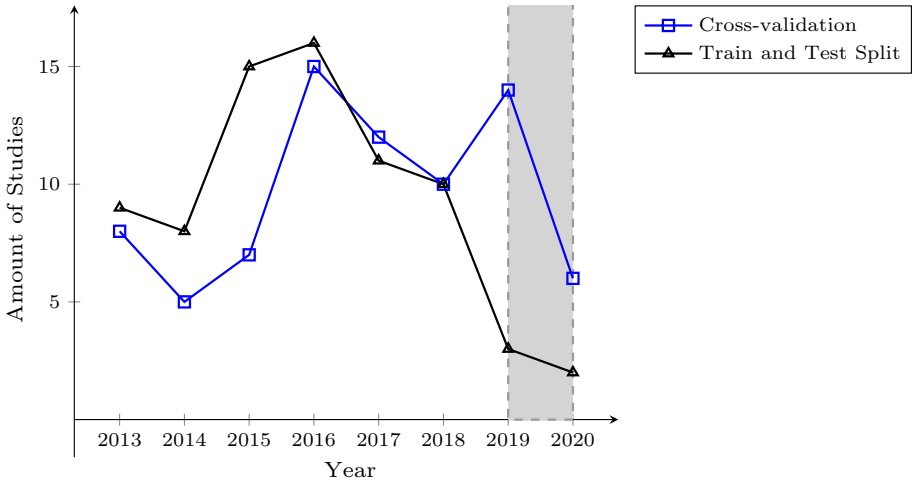

Fig. 8 Most used validation methods used to evaluate FS methods over the years. The gray box indicates that data collected for 2020 may be preliminary since this review was updated in October 2020

Table 13 Statistical significance tests used in studies to reject or not the null hypothesis

Statistical test	Studies	Percentage(%)
Does not perform any statistical test	123	70.29
Student's *t*-Test	31	17.71
Wilcoxon	7	4.00
Friedman-test	5	2.86
Nemenyi	3	1.71
Analysis of Variance (ANOVA)	2	1.14
Z-test	2	1.14
Chi-square	1	0.57
Cohen's Kappa Statistic	1	0.57
Total	175	100

7 Research trends and discussion

Based on the analysis of the problems, methods, and experiment settings raised in this review, we found relevant research trends and discussion points. In this section, we detail these research trends presenting our view on each of them. Sections 7.1– 7.4 present research trends and discussions based on each perspective of categorization model that we propose in Sect. 5. Sections 7.5 to 7.7 present research trends and discussions about experiment settings mapped in Sect. 6.

Table 14 Filter Strategies versus Wrapper strategies over the years

Year	Total of studies	Filter strategies		Wrapper strategies	
		Studies	Percentage	Studies	Percentage (%)
2013	22	20	90.91	2	9.09
2014	16	13	81.25	2	12.50
2015	28	23	82.14	5	17.86
2016	33	22	66.67	7	21.21
2017	27	20	74.07	7	25.93
2018	22	16	72.73	5	22.73
2019	18	8	44.44	9	50.00
2020	9	4	44.44	5	55.56
Total	175	114	65.14	39	22.29

7.1 Filter has been the feature selection dominant strategy for text classification, but a change is coming

In Sect. 5.1, we identified that most studies about FS for text classification implement the filter strategy. We found three main reasons for this preference for filter strategy (Kumar 2014; Tang et al. 2014):

- Simplicity—Filter-based methods usually have simpler design and development than wrapper and embedded methods.
- Classifier independence—In filter strategy, the result of FS is not biased by choice of classifiers.
- Computationally efficient—Filter-based methods are efficient and fast to compute. This advantage becomes even more important for text classification and other problems with high-dimensional data.

Despite filtering be the widely used strategy, we found that the percentage of studies based on filter strategy is decreasing, and the rate of studies based on wrapper strategy is increasing as shown in Table 14 and Fig. 9. The columns about Filter Strategy encompass the Two-Step (Filter+Filter) studies and the columns about Wrapper Strategy cover Hybrid (Filter+Wrapper) studies. We believe that the percentage of studies using other strategies (wrapper, embedded, and hybrid) will continues to increase. We see the following reasons for this increase:

- Large volume of published studies using the filter strategy – Since the volume of work using the filter strategy is large, we believe that the margin for improvement of results using this strategy is reduced. Therefore, we see that researchers tend to explore other strategies to pursue better results.
- Evolution of computing power and cost – The increase in processing power and computational cost reduction facilitates research techniques that are more computationally intensive, such as wrappers methods. In other words, the computational efficiency of the filter strategy tends to become a less important factor, as the available computing power increases.

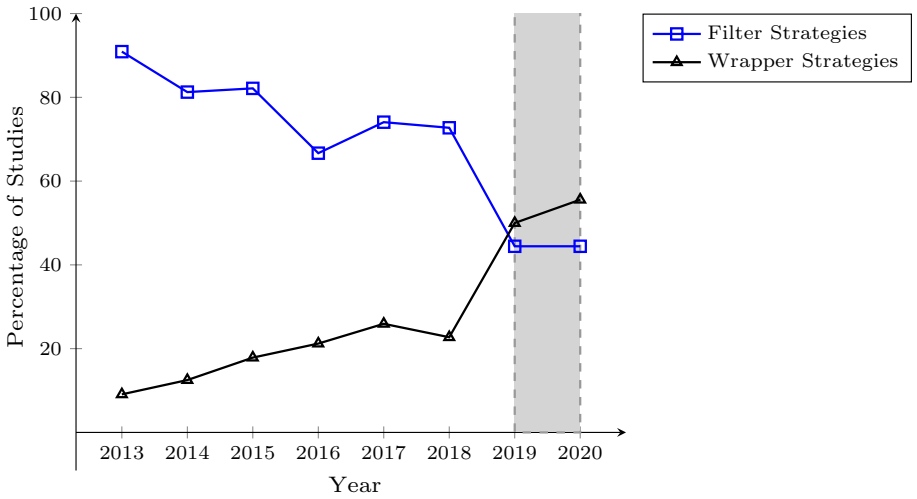

Fig. 9 Percentage of Filter Strategies vs Percentage of Wrapper Strategies over the years. The gray box indicates that data collected for 2020 may be preliminary since this review was updated in October 2020

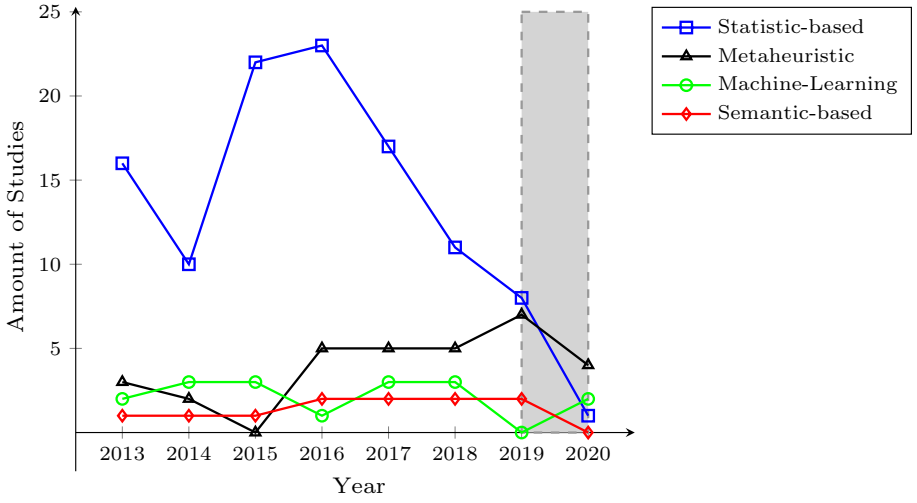

Fig. 10 Amount of FS studies by approach over the years. The gray box indicates that data collected for 2020 may be preliminary since this review was updated in October 2020

7.2 Metaheuristic approach is the trend

In Sect. 5.2, we identified that most studies about FS for text classification are mainly based on the statistic-based approach. However, we have analyzed the evolution of approaches used by grouping them by publication year (Fig. 10), and we noticed that the number of studies based on statistical approaches has been decreasing since 2016. On the same graph, we can see a gradual increase in the number of studies based on metaheuristics from 2015

to 2019. In 2016, 4 times more studies were published based on statistical approaches compared to the number of studies based on metaheuristics in the same year. On the other hand, almost the same number of studies were published in each approach during 2019. If this trend continues, in the coming years, the predominant approach will be metaheuristic.

We believe that the increasing use of the metaheuristic approach is motivated by the same factors then the use of the wrapper strategy (discussed in Sect. 7.1). Similarly, a considerable volume of studies is already available based on purely statistical approaches. In this way, researchers tend to have a smaller margin to achieve better results using the same approach. For this reason, they tend to explore more sophisticated approaches, such as metaheuristic techniques. The increasing use of the metaheuristic approach can also be directly related to the wrapper strategy's greater use. Wrapper strategy is frequently employed to search for the best subset of features (as explained in Sect. 4). Since this subset search is a hard problem (as explained in Sect. 4.2), metaheuristic search techniques are usually the solution adopted.

In addition to the two predominant approaches (statistical and metaheuristic), semantic-based and machine learning-based approaches have also been used in a relevant number of studies. However, as it is possible to observe in the graph by year (Fig. 10), neither approach had a significant increase or decrease in the volume of studies per year. Approaches that have had few scattered studies over the years were not included in this chart. For example, we only found one study mainly based on the grammatical approach published in 2015 and three rule-based studies published in 2016, 2018, and 2019.

Considering the observations made in the previous paragraphs, we conclude that the metaheuristic approach tends to become prevalent in the coming years and the number of studies based mainly on a statistical approach tends to continue decreasing. We also believe that researchers will tend to combine two or more approaches in the same study to seek better results. This review focused on mapping the principal approach used in each study. A future work could be examining all secondary approaches used in each study and how they are being combined.

7.3 Multiclass classifiers are still dominant

In Sect. 5.3, we identified that most studies were evaluated or designed to multiclass classifiers (67.43% of total) and binary classifiers (24.57% of total). To assess the possible change of this trend, we analyzed the distribution by year (Fig. 11). Disregarding the year 2020 with preliminary data, it is not possible to identify any clear trend indicating a change in this distribution. Despite this, we believe that the number of multi-label studies will increase due to the popularity increase of multi-label classification (Pereira et al. 2018). However, we believe that this type of change tends to be gradual. The reason is that new FS studies tend to use the same types of classifiers and datasets that have been widely used in previous studies to facilitate comparison between studies.

7.4 Supervised versus unsupervised feature selection methods

In Sect. 5.4, we identified that most studies are based on supervised techniques (94.86% of total). To assess if there is any sign of growth in studies of unsupervised or semi-supervised techniques, we analyzed this distribution by year (Table 15). However, from this table, we can see that the largest number of unsupervised and semi-supervised studies were

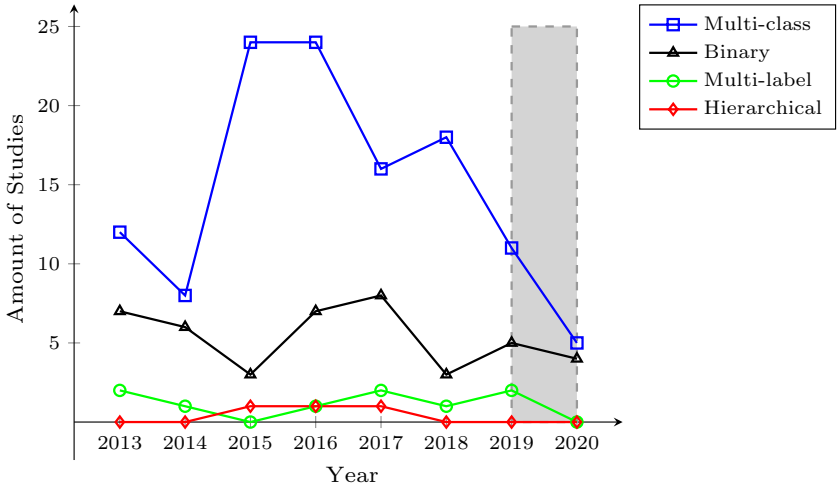

Fig. 11 Amount of FS studies by type of classification over the years. The gray box indicates that data collected for 2020 may be preliminary since this review was updated in October 2020

Table 15 Number of supervised, unsupervised and semi-supervised studies over the years

Year	Supervised	Unsupervised	Semi-supervised	Percentage of supervised studies (%)
2013	21	0	1	95.45
2014	15	0	1	93.75
2015	28	0	0	100.00
2016	29	3	1	87.88
2017	25	1	1	92.59
2018	21	1	0	95.45
2019	18	0	0	100.00
2020	9	0	0	100.00
Total	166	5	4	94.86

concentrated in 2017 without progression in the following years. Thus, we believe that for the next few years, studies will probably remain focused on supervised techniques.

7.5 Recent researches still over old public datasets: the need for new benchmarks

In Sect. 6.2, we present that the three most commonly used datasets in the studies are Reuters-21578, 20NewsGroup, and WebKb. Table 16 shows the year of creation and the size (number of documents) each of these datasets. Note that the three datasets are over 20 years old and have a volume below 22,000 documents. These datasets can be considered old and small compared to other datasets like Reuters Corpus Volume I (RCV1). It is a dataset of over 800,000 manually categorized newswire stories made available by Reuters Ltd. for research purposes (Lewis et al. 2004). Although the RCV1 be a well-known benchmark for text classification with more than 2,000 studies citing the original paper (Lewis

Table 16 Age and size of most used datasets in experiments

Dataset	Year of creation	Number of documents	Reference
Reuters-21578	1987	21.578	Lewis (2019)
20NewsGroup	1995	20.000	Rennie (2019)
WebKB	1997	8.282	Webkb (2019)

Table 17 Historical trends in the usage of content languages for websites (W3Techs 2019)

Position	Language	Percentage (%)	Position	Language	Percentage (%)
1	English	54.40	11	Polish	1.60
2	Russian	6.70	12	Chinese	1.60
3	German	5.30	13	Dutch, Flemish	1.10
4	Spanish	4.90	14	Korean	0.90
5	French	3.70	15	Czech	0.90
6	Japanese	3.40	16	Vietnamese	0.80
7	Portuguese	2.70	17	Arabic	0.70
8	Italian	2.10	18	Greek	0.60
9	Persian	2.10	19	Hungarian	0.50
10	Turkish	1.60	20	Swedish	0.50

et al. 2004), none of the FS studies included in this review employed this dataset in their experiments.

We believe that most current studies still use those same datasets to facilitate a comparison of its results to previous studies. Thus, the use of these datasets tends to be preserved. One way to solve this problem is by using both classical as well as newer/larger datasets in new studies. In this way, it will be possible to compare the results to previous works and evaluate the methods using larger benchmarks.

7.6 The english language dominance

As explained in Sect. 6.2, most studies (72.57%) evaluate techniques only in English datasets. We believe that one of the reasons for this focus on the English language is that about 54.40% of internet web pages is written in this language (W3Techs 2019). The remainder of the web pages is distributed over several other languages, such as Russian, German, and Spanish. However, each one of these languages owns less than 7% of the webpages each (W3Techs 2019). Table 17 lists the 20 most widely used languages on the internet as of September 2019.

7.7 Feature selection is already a mature field allowing statistical evaluations

As presented in Sect. 6.4, most studies (70.29%) do not perform statistical significance tests to reject or not the null hypothesis. Analyzing the data grouping by year (Table 18),

Table 18 Statistical significance tests used in studies to reject or not the null hypothesis

Year	Studies	Using significance test	Percentage (%)
2013	22	3	13.64
2014	16	2	12.50
2015	28	8	28.57
2016	33	9	27.27
2017	27	7	25.93
2018	22	7	31.82
2019	18	10	55.56
2020	9	6	66.67
Total	175	52	29.71

Table 19 Statistical significance tests used in studies to reject or not the null hypothesis (Conference Studies versus Journal Studies)

Year	Conference studies			Journal studies		
	Studies	Using statistical tests	Percentage (%)	Studies	Using statistical tests	Percentage (%)
2013	14	0	0.00	8	3	37.50
2014	10	0	0.00	6	2	33.33
2015	18	3	16.67	10	5	50.00
2016	18	3	16.67	15	6	40.00
2017	13	2	15.38	14	5	35.71
2018	9	1	11.11	13	6	46.15
2019	8	2	25.00	10	8	80.00
2020	0	0	0.00	9	6	66.67
Total	90	11	12.22	85	41	48.24

we noticed a progressive increase in the use of statistical tests. We believe that this increase indicates a maturing of the research area.

Analyzing publications in conferences and journals separately (Table 19), we concluded that the use of statistical tests is widespread in papers published in journals. From 90 articles published in conference, only 11 studies (12.22%) use statistical tests to support their findings. On the other hand, 41 of 85 studies (48.24%) published in journals use statistical tests. In both cases, there is an increase in the use of statistical tests over the years. We believe that this demonstrates an increase in maturity in FS studies for text classification.

8 Conclusion

The volume and diversity of studies included in this SLR show the complexity and importance of FS methods tailored for text classification problems. The categorization scheme for FS methods that we propose in this article is an artifact that allows an

analysis of the current state of research and also the positioning of new studies. The proposed categorization scheme enables grouping studies into different perspectives so that it is possible to observe the similarities and differences among the methods. Besides, another contribution of this SLR is a mapping of experiment settings. We hope that this mapping will help the design of new experiments by presenting what settings are most commonly used and what settings that are still unexplored. Finally, we presented a discussion about the main findings, and we pointed out the main trends and gaps identified by our SLR. The main future work that we have identified is the development is an open platform for comparing and testing FS methods specific for text classification.

Appendix A. List of acronyms

ACC	Accuracy Measure
ACC2	Balanced Accuracy Measure
ALOFT	At Least One FeaTure
ANOVA	Analysis of Variance
ACA	Bit-priori Association Classification Algorithm
BBHA	Binary Black Hole Algorithm
BFSM	Blended Feature Selection Method
BGSA	Binary Gravitational Search Algorithm
BMI	Balanced Mutual Information
BoDW	Bag of Discriminative Words
BoW	Bag of Words
BPSO	Binary Particle Swarm Optimization
CAS	Correlative Association Score
CDM	Class Discriminating Measure
CHI	Chi-square
CMFS	Comprehensively Measure Feature Selection
CNN	Convolutional Neural Network
CrowdFS	Crowd-based Feature Selection
CSO	Cat Swarm Optimization
DBN	Deep Belief Network
DF	Document Frequency
DFS	Discriminative Features Selection
DFS	Distinguishing Feature Selector
DGBFS	Diversified Greedy Backward-Forward Search
DPP	Discriminative Personal Purity
DT	Decision Tree
EEFS	Ensemble Embedded Feature Selection
FRFS	Fuzzy Rough Feature Selection
FS	Feature Selection
GAWA	Genetic Algorithm and Wrapper Approaches
GFSS	Global Filter-based Feature Selection Scheme
GI	Gini Index
GPSO	Geometric Particle Swarm Optimization
HAN	Hierarchical Attention Network

HRFS	Hebb Rule Based Feature Selection
IDF	Inverse Document Frequency
IG	Information Gain
IPSO	Improved Particle Swarm Optimization
ISCA	Improved Sine Cosine Algorithm
KNN	*k*-Nearest Neighbors
LDA	Latent Dirichlet Allocation
LSAN	Latent Selection Augmented Naive Bayes
MBF	Markov Blanket Filter
MFS	Meta Feature Selection
MFSLFD	Memetic Feature Selection based on Label Frequency Difference
MI	Mutual Information
MMI	Multivariate Mutual Information
MMR	Max-Min Ratio
MOANOFS	Multi-Objective Automated Negotiation based Online Feature Selection
MORDC	Multi-Objective Relative Discriminative Criterion
MRDC	Multivariate Relative Discrimination Criterion
NB	Naive Bayes
NDM	Normalized Difference Measure
OR	Odds Ratio
OS-FS	Optimized Swarm Search-based Feature Selection
PCT	Pairwise Comparison Transformation
POS	Part of Speech
POSFilter	Part of Speech Filter
PSO	Particle Swarm Optimization
RCV1	Reuters Corpus Volume I
RDC	Relative Discrimination Criterion
RF	Random Forest
RFE	Recursive Feature Elimination
RP-GSO	Random Projection and Gram-Schmidt Orthogonalization
SAIG	Sparsity Adjusted Information Gain
SBATFS	Spark BAT Feature Selection
SIGCHI	Square of Information Gain and Chi-square
SLR	Systematic Literature Review
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machines
SVM-RFE	Support Vector Machine-Recursive Feature Elimination
SWA	Small World Algorithm
t-Test	Student's *t*-Test
TF	Term Frequency
TF-IDF	Term Frequency-Inverse Document Frequency
WFSAIG	Wrapper Feature Selection Algorithm based on Iterated Greedy
WI-OMFS	Wolf Intelligence Based Optimization of Multi-Dimensional Feature Selection Approach

References

Abdollahi M, Gao X, Mei Y, Ghosh S, Li J (2019) An ontology-based two-stage approach to medical text classification with feature selection by particle swarm optimisation. In: Proceedings of the IEEE congress on evolutionary computation, pp 119–126

Agnihotri D, Verma K, Tripathi P (2017) Variable global feature selection scheme for automatic classification of text documents. Expert Syst Appl 81:268–281. https://doi.org/10.1016/j.eswa.2017.03.057

Agnihotri D, Verma K, Tripathi P (2016) Computing correlative association of terms for automatic classification of text documents. Proceedings of the international symposium on computer vision and the internet, https://doi.org/10.1145/2983402.2983424

Agnihotri D, Verma K, Tripathi P (2017a) Mutual information using sample variance for text feature selection. In: Proceedings of the international conference on communication and information processing, pp 39–44, https://doi.org/10.1145/3162957.3163054

Agnihotri D, Verma K, Tripathi P, Singh B (2018) Soft voting technique to improve the performance of global filter based feature selection in text corpus. Appl Intell 49. https://doi.org/10.1007/s10489-018-1349-1

Agun HV, Yilmazel O (2019) Incorporating topic information in a global feature selection schema for authorship attribution. IEEE Access 7:98522–98529

Al-Salemi B, Ayob M, Noah SAM (2018) Feature ranking for enhancing boosting-based multi-label text categorization. Expert Syst Appl 113:531–543. https://doi.org/10.1016/j.eswa.2018.07.024

Al-Salemi B, Ayob M, Noah SAM, Aziz MJA (2017) Feature selection based on supervised topic modeling for boosting-based multi-label text categorization. In: Proceedings of the international conference on electrical engineering and informatics, pp 1–6, https://doi.org/10.1109/ICEEI.2017.8312411

Alshalabi H, Tiun S, Omar N, Albared M (2013) Experiments on the use of feature selection and machine learning methods in automatic Malay text categorization. Procedia Technol 11:748–754. https://doi.org/10.1016/J.PROTCY.2013.12.254

Arani SHS, Mozaffari S (2013) Genetic-based feature selection for spam detection. In: Proceedings of the Iranian conference on electrical engineering, https://doi.org/10.1109/IranianCEE.2013.6599551

Baccianella S, Esuli A, Sebastiani F (2013) Using micro-documents for feature selection: the case of ordinal text classification. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2013.02.010

Baccianella S, Esuli A, Sebastiani F (2014) Feature selection for ordinal text classification. Neural Comput

Badawi D, Altincay H (2014) A novel framework for termset selection and weighting in binary text classification. Eng Appl Artif Intell 35:38–53. https://doi.org/10.1016/j.engappai.2014.06.012

Baggenstoss PM (2003) The PDF projection theorem and the class-specific method. IEEE Trans Sig Process 51(3):672–685. https://doi.org/10.1109/TSP.2002.808109

Bagheri A, Saraee M, De Jong F (2013) Sentiment classification in Persian: introducing a mutual information-based method for feature selection. In: Proceedings of the Iranian conference on electrical engineering, https://doi.org/10.1109/IranianCEE.2013.6599671

Bahassine S, Madani A, Al-Sarem M, Kissi M (2018) Feature selection using an improved Chi-square for Arabic text classification. J King Saud Univ—Comput Inf Sci. https://doi.org/10.1016/j.jksuci.2018.05.010

Bahassine S, Madani A, Kissi M (2016) An improved Chi-square feature selection for Arabic text classification using decision tree. In: Proceedings of the international conference on intelligent systems: theories and applications, pp 1–5, https://doi.org/10.1109/SITA.2016.7772289

Bai X, Gao X, Xue B (2018) Particle swarm optimization based two-stage feature selection in text mining. In: Proceedings of the IEEE congress on evolutionary computation, pp 1–8

Belazzoug M, Touahria M, Nouioua F, Brahimi M (2020) An improved sine cosine algorithm to select features for text categorization. J King Saud Univ—Comput Inf Sci 32(4):454–464. https://doi.org/10.1016/j.jksuci.2019.07.003

Benitez IP, Sison AM, Medina RP (2018) An improved genetic algorithm for feature selection in the classification of disaster-related Twitter messages. In: Proceedings of the IEEE symposium on computer applications and industrial electronics, https://doi.org/10.1109/ISCAIE.2018.8405477

BenSaid F, Alimi AM (2021) Online feature selection system for big data classification based on multi-objective automated negotiation. Pattern Recognit 110:107629. https://doi.org/10.1016/j.patcog.2020.107629

Bergstra J, Bengio Y (2013) Random search for hyper-parameter optimization. J Mach Learn Res

Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguist 5:135–146

Braytee A, Liu W, Catchpoole D, Kennedy P (2017) Multi-label feature selection using correlation information. In: Proceedings of the ACM on conference on information and knowledge management, pp 1649–1656, https://doi.org/10.1145/3132847.3132858

Canuto S, Sousa DX, Gonçalves MA, Rosa TC (2018) A thorough evaluation of distance-based meta-features for automated text classification. IEEE Trans Knowl Data Eng 11(10):346–347. https://doi.org/10.1007/s10489-018-1349-10

Cekik R, Uysal AK (2020) A novel filter feature selection method using rough set for short text data. Expert Syst Appl 160:113691. https://doi.org/10.1007/s10489-018-1349-11

Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40(1):16–28. https://doi.org/10.1007/s10489-018-1349-12

Chen H, Hou Q, Han L, Hu Z, Ye Z, Zeng J, Yuan J (2019) Distributed text feature selection based on bat algorithm optimization. Proc IEEE Int Conf Intell Data Acquis Adv Comput Syst Technol Appl 1:75–80

Chen Y, Han B, Hou P (2014) New feature selection methods based on context similarity for text categorization. In: Proceedings of the international conference on fuzzy systems and knowledge discovery, https://doi.org/10.1109/FSKD.2014.6980902

Chen H, Hou Y, Luo Q, Hu Z, Yan L (2018) Text feature selection based on water wave optimization algorithm. In: Proceedings of the international conference on advanced computational intelligence, https://doi.org/10.1109/ICACI.2018.8377518

Chen L, Li J, Zhang L (2017) A method of text categorization based on genetic algorithm and LDA. In: Proceedings of the chinese control conference, https://doi.org/10.23919/ChiCC.2017.8029089

Chen X, Ma J, Lu Y (2013) Feature selection for Chinese online reviews sentiment classification. In: Proceedings of the joint conference of international conference on computational problem-solving and international high speed intelligent communication forum, https://doi.org/10.1109/ICCPS.2013.6893490

Chopard B, Tomassini M (2018) An introduction to metaheuristics for optimization. Springer Int Publ. https://doi.org/10.1007/978-3-319-93073-2

Chormunge S, Jena S (2018) Correlation based feature selection with clustering for high dimensional data. J Electl Syst Inf Technol 5(3):542–549. https://doi.org/10.1016/J.JESIT.2017.06.004

Demsar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30

Deng X, Li Y, Weng J, Zhang J (2019) Feature selection for text classification: a review. Multimed Tools Appl 78(3):3797–3816. https://doi.org/10.1007/s11042-018-6083-5

Ekbal A, Saha S (2015) Joint model for feature selection and parameter optimization coupled with classifier ensemble in chemical mention recognition. Knowl-Based Syst. https://doi.org/10.1016/j.knosys.2015.04.015

Feng G, Guo J, Jing BY, Sun T (2015a) Feature subset selection using naive Bayes for text classification. Pattern Recognit Lett. https://doi.org/10.1016/j.patrec.2015.07.028

Feng L, Zuo W, Wang Y (2015b) Improved comprehensive measurement feature selection method for text categorization. In: Proceedings of the international conference on network and information systems for computers, https://doi.org/10.1109/ICNISC.2015.34

Ferilli S, De Carolis B, Esposito F, Redavid D (2015) Sentiment analysis as a text categorization task: a study on feature and algorithm selection for Italian language. In: Proceedings of the IEEE international conference on data science and advanced analytics, https://doi.org/10.1109/DSAA.2015.7344882

Ferreira CHP, De Medeiros DMR, Santana F (2016) FCFilter: feature selection based on clustering and genetic algorithms. In: Proceedings of the IEEE congress on evolutionary computation, https://doi.org/10.1109/CEC.2016.7744048

Fong S, Gao E, Wong R (2016) Optimized swarm search-based feature selection for text mining in sentiment analysis. In: Proceedings of the IEEE international conference on data mining workshop, pp 1153–1162, https://doi.org/10.1109/ICDMW.2015.231

Forman G (2004) A pitfall and solution in multi-class feature selection for text classification. Proceed Int Conf Mach Learn 10(1145/1015330):1015356

Fragoso RCP, Pinheiro RHW, Cavalcanti GDC (2016) Class-dependent feature selection algorithm for text categorization. In: Proceedings of the international joint conference on neural networks, vol 2016-Octob, https://doi.org/10.1109/IJCNN.2016.7727649

Fragoso RCP, Pinheiro RHW, Cavalcanti GDC (2017) A method for automatic determination of the feature vector size for text categorization. In: Proceedings of the Brazilian conference on intelligent systems, https://doi.org/10.1109/BRACIS.2016.055

Fukumoto F, Suzuki Y (2015) Temporal-based feature selection and transfer learning for text categorization. In: Proceedings of the international joint conference on knowledge discovery, knowledge engineering and knowledge management, http://socrates.acadiau.ca/courses/comp/dsilver/

Gao Z, Xu Y, Meng F, Qi F, Lin Z (2014) Improved information gain-based feature selection for text categorization. In: Proceedings of the international conference on wireless communications, vehicular technology, information theory and aerospace and electronic systems

Ghareb AS, Bakar AA, Hamdan AR (2016) Hybrid feature selection based on enhanced genetic algorithm for text categorization. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2015.12.004

Ghareb AS, Abu Bakara A, Al-Radaideh QA, Hamdan AR (2018) Enhanced filter feature selection methods for Arabic text categorization. Int J Inf Retr Res. https://doi.org/10.4018/IJIRR.2018040101

Gökalp O, Tasci E, Ugur A (2020) A novel wrapper feature selection algorithm based on iterated greedy metaheuristic for sentiment classification. Expert Syst Appl 146:113176. https://doi.org/10.1016/j.eswa.2020.113176

Gunduz H, Cataltepe Z (2015) Borsa Istanbul (BIST) daily prediction using financial news and balanced feature selection. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2015.07.058

Guo Y, Chung F, Li G, Zhang L (2019) Multi-label bioinformatics data classification with ensemble embedded feature selection. IEEE Access 7:103863–103875

Guo Y, Chung F, Li G (2017) An ensemble embedded feature selection method for multi-label clinical text classification. In: Proceedings of the IEEE international conference on bioinformatics and biomedicine, https://doi.org/10.1109/BIBM.2016.7822631

Guru DS, Ali M, Suhil M (2018) A novel term weighting scheme and an approach for classification of agricultural arabic text complaints. In: Proceedings of the IEEE international workshop on arabic and derived script analysis and recognition, pp 24–28

Guru DS, Suhil M, Raju LN, Kumar NV (2018) An alternative framework for univariate filter based feature selection for text categorization. Pattern Recog Lett 103(2018):23–31. https://doi.org/10.1016/j.patrec.2017.12.025

Guru D, Swarnalatha K, Kumar VN, Anami B (2020) Effective technique to reduce the dimension of text data. Int J Comput Vis Image Process 10:67–85. https://doi.org/10.4018/IJCVIP.2020010104

Hagenau M, Liebmann M, Neumann D (2013) Automated news reading: stock price prediction based on financial news using context-capturing features. Decis Support Syst 55(3):685–697. https://doi.org/10.1016/j.dss.2013.02.006

Hai NT, Nghia NH, Le TD, Nguyen VT (2015) A hybrid feature selection method for Vietnamese text classification. In: Proceedings of the IEEE international conference on knowledge and systems engineering, https://doi.org/10.1109/KSE.2015.25

Han J, Zuo W, Liu L, Xu Y, Peng T (2016) Building text classifiers using positive, unlabeled and 'outdated' examples. Concurr Comput. https://doi.org/10.1002/cpe.3879

Higgins JPT, Green S (2008) Cochrane handbook for systematic reviews of interventions: cochrane book series. Wiley, New York. https://doi.org/10.1002/9780470712184

Hussain S, Keung J, Khan AA (2017) Software design patterns classification and selection using text categorization approach. Appl Soft Comput 58:225–244. https://doi.org/10.1016/J.ASOC.2017.04.043

Hussain SF, Babar HZUD, Khalil A, Jillani RM, Hanif M, Khurshid K (2020) A fast non-redundant feature selection technique for text data. IEEE Access 8:181763–181781. https://doi.org/10.1109/ACCESS.2020.3028469

Imani MB, Keyvanpour MR (2013) Azmi R (2013) A novel embedded feature selection method: a comparative study in the application of text categorization. Appl Artif Intell 10(1080/08839514):774211

Islam M, Anjum A, Ahsan T, Wang L (2019) Dimensionality reduction for sentiment classification using machine learning classifiers. In: Proceedings of the IEEE symposium series on computational intelligence, pp 3097–3103

Japkowicz N (2000) The class imbalance problem: significance and strategies. In: Proceedings of the international conference on artificial intelligence

Javed K, Maruf S, Babri HA (2015) A two-stage Markov blanket based feature selection algorithm for text classification. Neurocomputing. https://doi.org/10.1016/j.neucom.2015.01.031

Jiang XY, Jin S (2013) An improved mutual information-based feature selection algorithm for text classification. In: Proceedings of the international conference on intelligent human-machine systems and cybernetics, https://doi.org/10.1109/IHMSC.2013.37

Jiang T, Yu H (2015) A novel feature selection based on Tibetan grammar for Tibetan text classification. In: Proceedings of the IEEE international conference on software engineering and service sciences, https://doi.org/10.1109/ICSESS.2015.7339093

Jie Y, Keping L (2019) The fault diagnosis model for railway system based on an improved feature selection method. In: Proceedings of the IEEE international conference on electronics information and emergency communication, pp 1–4

Karabulut M (2013) Fuzzy unordered rule induction algorithm in text categorization on top of geometric particle swarm optimization term selection. Knowl Based Syst 54:288–297. https://doi.org/10.1016/J.KNOSYS.2013.09.020

Kermani FZ, Eslami E, Sadeghi F (2019) Global filter-wrapper method based on class-dependent correlation for text classification. Eng Appl Artif Intell 85:619–633. https://doi.org/10.1016/j.engappai.2019.07.003

Kim K, Zzang S (2018) Trigonometric comparison measure: a feature selection method for text categorization. Data Knowl Eng 119. https://doi.org/10.1016/j.datak.2018.10.003

Kitchenham B (2004) Procedures for performing systematic reviews. Tech. Rep. TR/SE-0401, Department of Computer Science, Keele University and National ICT

Kowsari K, Jafari Meimandi K, Heidarysafa M, Mendu S, Barnes L, Brown D (2019) Text classification algorithms: a survey. Inf Switz 10. https://doi.org/10.3390/info10040150

Kumar HMK, Harish BS (2018) Sarcasm classification: a novel approach by using content based feature selection method. Procedia computer science 143:378–386. https://doi.org/10.1016/j.procs.2018.10.409, 8th international conference on advances in computing and communications (ICACC-2018)

Kumar V (2014) Feature selection a literature review. Smart Comput Rev. https://doi.org/10.6029/smartcr.2014.03.007

Kumbhar P, Mali M (2013) A survey on feature selection techniques and classification algorithms for efficient text slassification. Int J Sci Res 14(5):2319–7064

Kumbhar P, Mali M, Atique M (2017) A genetic-fuzzy approach for automatic text categorization. In: Proceedings of the international advance computing conference, https://doi.org/10.1109/IACC.2017.114

Kun YJ, Lei Z (2014) Sentiment feature selection algorithm for Chinese micro-blog. In: Proceedings of the international conference on management of e-commerce and e-government, pp 114–118, https://doi.org/10.1109/ICMeCG.2014.32

Kyaw KS, Limsiroratana S (2019) Towards nature-inspired intelligence search for optimization of multi-dimensional feature selection. In: Proceedings of the international computer science and engineering conference, pp 379–384

Labani M, Moradi P, Ahmadizar F, Jalili M (2018) A novel multivariate filter method for feature selection in text classification problems. Eng Appl Artif Intell 70(November 2016):25–37. https://doi.org/10.1016/j.engappai.2017.12.014

Labani M, Moradi P, Jalili M (2020) A multi-objective genetic algorithm for text feature selection using the relative discriminative criterion. Expert Syst Appl 149:113276. https://doi.org/10.1016/j.eswa.2020.113276

Lampos V, Zou B, Cox IJ (2017) Enhancing feature selection using word embeddings. Proc Int Conf World Wide Web 10(1145/3038912):3052622

Lan Y, Hao Y, Xia K, Qian B, Li C (2020) Stacked residual recurrent neural networks with cross-layer attention for text classification. IEEE Access 8:70401–70410

Larabi Marie-Sainte S, Alalyani N (2018) Firefly algorithm based feature selection for Arabic text classification. J King Saud Univ Comput Inf Sci. https://doi.org/10.1016/J.JKSUCI.2018.06.004

Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C, De Schaetzen V, Duque R, Bersini H, Nowé A (2012) A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Trans Comput Biol Bioinf. https://doi.org/10.1109/TCBB.2012.33

Lee J, Kim DW (2013) Feature selection for multi-label classification using multivariate mutual information. Pattern Recognit Lett. https://doi.org/10.1016/j.patrec.2012.10.005

Lee J, Kim DW (2015) Mutual information-based multi-label feature selection using interaction information. Expert Syst Appl 42(4):2013–2025. https://doi.org/10.1016/j.eswa.2014.09.063

Lee J, Yu I, Park J, Kim DW (2019) Memetic feature selection for multilabel text categorization using label frequency difference. Inf Sci 485:263–280. https://doi.org/10.1016/j.ins.2019.02.021

Lewis DD (2019) Reuters-21578 text categorization collection data set. https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection

Lewis DD, Yang Y, Rose TG, Li F (2004) RCV1: a new benchmark collection for text categorization research. J Mach Learn Res 5:361–397

Li B (2016a) Importance weighted feature selection strategy for text classification. In: Proceedings of the international conference on Asian language processing

Li B (2016b) Selecting features with class based and importance weighted document frequency in text classification. In: Proceedings of the ACM symposium on document engineering, pp 139–142, https://doi.org/10.1145/2960811.2967164
Li J (2013) An approach to meta feature selection. In: Proceedings of the Canadian conference on electrical and computer engineering, https://doi.org/10.1109/CCECE.2013.6567849
Li Z, Lu W, Sun Z, Xing W (2016) A parallel feature selection method study for text classification. Neural Comput Appl 28:1–12. https://doi.org/10.6029/smartcr.2014.03.0070
Liang J, Zhou X, Guo L, Bai S (2015) Feature selection for sentiment classification using matrix factorization. In: Proceedings of the international conference on world wide web, pp 63–64, https://doi.org/10.1145/2740908.2742741
Lifang Y, Sijun Q, Huan Z (2017) Feature selection algorithm for hierarchical text classification using Kullback-Leibler divergence. In: Proceedings of the ieee international conference on cloud computing and big data analysis, https://doi.org/10.1109/ICCCBDA.2017.7951950
Li Q, He L, Lin X (2013a) Categorical term frequency probability based feature selection for document categorization. In: Proceedings of the international conference on soft computing and pattern recognition, https://doi.org/10.1109/SOCPAR.2013.7054103
Li Q, He L, Lin X (2013b) Dimension reduction based on categorical fuzzy correlation degree for document categorization. In: Proceedings of the IEEE international conference on granular computing, https://doi.org/10.1109/GrC.2013.6740405
Li Q, He L, Lin X (2014) Improved categorical distribution difference feature selection for Chinese document categorization. In: Proceedings of the international conference on ubiquitous information management and communication
Li L, Li C (2015) Research and improvement of a spam filter based on naive Bayes. In: Proceedings of the international conference on intelligent human-machine systems and cybernetics, https://doi.org/10.1109/IHMSC.2015.208
Lin KC, Zhang KY, Huang YH, Hung JC, Yen N (2016) Feature selection based on an improved cat swarm optimization algorithm for big data classification. J Supercomput. https://doi.org/10.6029/smartcr.2014.03.0071
Liu Y, Wang Y, Feng L, Zhu X (2016) Term frequency combined hybrid feature selection method for spam filtering. Pattern Anal Appl. https://doi.org/10.6029/smartcr.2014.03.0072
Li B, Yan Q, Xu Z, Wang G (2015) Weighted document frequency for feature selection in text classification. In: Proceedings of international conference on Asian language processing, https://doi.org/10.1109/IALP.2015.7451549
Li J, Zhao J, Lu K (2016a) Joint feature selection and structure preservation for domain adaptation. In: Proceedings of the IJCAI international joint conference on artificial intelligence
Lu Y, Chen Y (2017) A text feature selection method based on the small world algorithm. Procedia Comput Sci 107:276–284. https://doi.org/10.6029/smartcr.2014.03.0073
Lu Y, Liang M, Ye Z, Cao L (2015) Improved particle swarm optimization algorithm and its application in text feature selection. Appl Soft Comput J. https://doi.org/10.6029/smartcr.2014.03.0074
Malji P, Sakhare S (2017) Significance of entropy correlation coefficient over symmetric uncertainty on FAST clustering feature selection algorithm. In: Proceedings of international conference on intelligent systems and control, https://doi.org/10.1109/ISCO.2017.7856035
Manning CD, Schutze H, Raghavan P (2008) Introduction to information retrieval. Cambridge University Press, Cambridge
Manochandar S, Punniyamoorthy M (2018) Scaling feature selection method for enhancing the classification performance of support vector machines in text mining. Comput Ind Eng 124:139–156. https://doi.org/10.1016/j.cie.2018.07.008
Mendez JR, Nez TRCY, Ruano-Ordas D (2019) A new semantic-based feature selection method for spam filtering. Appl Soft Comput 76:89–104. https://doi.org/10.1016/j.asoc.2018.12.008
Mikolov T, Chen K, Corrado G, Dean J (2013a) Efficient estimation of word representations in vector space. In: Proceedings of the international conference on learning representations - workshop track proceedings
Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013b) Distributed representations of words and phrases and their compositionality. In: Proceedings of the advances in neural information processing systems, pp 3111–3119
Mirończuk MM, Protasiewicz J (2018) A recent overview of the state-of-the-art elements of text classification. Expert Syst Appl 106:36–54. https://doi.org/10.1016/j.eswa.2018.03.058
Mladenović M, Mitrović J, Krstev C, Vitas D (2016) Hybrid sentiment analysis framework for a morphologically rich language. J Intell Inf Syst. https://doi.org/10.1007/s10844-015-0372-5

Moher D, Liberati A, Tetzlaff J, Altman DG (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. J Clin Epidemiol. https://doi.org/10.1016/j.jclinepi.2009.06.005

Nag K, Pal NR (2016) A multiobjective genetic programming-based ensemble for simultaneous feature selection and classification. IEEE Trans Cybernet. https://doi.org/10.1109/TCYB.2015.2404806

Naik A, Rangwala H (2016) Embedding feature selection for large-scale hierarchical classification. In: Proceedings of the IEEE international conference on big data, https://doi.org/10.1109/BigData.2016.7840725

Nam LNH, Quoc HB (2016) A combined approach for filter feature selection in document classification. In: Proceedings of the international conference on tools with artificial intelligence, https://doi.org/10.1109/ICTAI.2015.56

Nogueira Rios T, Gama Bispo BV (2018) Statera: a balanced feature selection method for text classification. In: Proceedings of the Brazilian conference on intelligent systems, pp 260–265

Onan A, Korukoglu S (2017) A feature selection model based on genetic rank aggregation for text sentiment classification. J Inf Sci 43(1):25–38. https://doi.org/10.1177/0165551515613226

Ong BY, Goh SW, Xu C (2015) Sparsity adjusted information gain for feature selection in sentiment analysis. In: Proceedings of the IEEE international conference on big data, pp 2122–2128, https://doi.org/10.1109/BigData.2015.7363995

Ortega-Mendoza RM, López-Monroy AP, Franco-Arcega A, Montes-y Gómez M (2018) Emphasizing personal information for author profiling: new approaches for term selection and weighting. Knowl Based Syst 145:169–181. https://doi.org/10.1016/J.KNOSYS.2018.01.014

Ouhbi B, Kamoune M, Frikh B, Zemmouri EM, Behja H (2016) A hybrid feature selection rule measure and its application to systematic review. In: Proceedings of the international conference on information integration and web-based applications and services, pp 106–114, https://doi.org/10.1145/3011141.3011177

Parlar T, Ozel SA, Song F (2016) A new feature selection method for sentiment analysis of Turkish reviews. In: Proceedings of the international symposium on innovations in intelligent systems and applications, pp 1–6, https://doi.org/10.1109/INISTA.2016.7571833

Pashaei E, Aydin N (2017) Binary black hole algorithm for feature selection and classification on biological data. Appl Soft Comput J 56:94–106. https://doi.org/10.1016/j.asoc.2017.03.002

Patil LH, Atique M (2013) A novel feature selection based on information gain using WordNet. In: Proceedings of the science and information conference

Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the conference on empirical methods in natural language processing, pp 1532–1543

Pereira RB, Plastino A, Zadrozny B, Merschmann LHC (2018) Categorizing feature selection methods for multi-label classification. Artif Intell Rev 49(1):57–78. https://doi.org/10.1007/s10462-016-9516-4

Pinheiro RHW, Cavalcanti GDC, Ren TI (2015) Data-driven global-ranking local feature selection methods for text categorization. Expert Syst Appl. https://doi.org/10.1016/j.asoc.2018.12.0080

Pintas JT, Correia L, Bicharra Garcia AC (2017) Crowd-based feature selection for document retrieval in highly demanding decision-making scenarios. Procedia Comput Sci 112:822–832. https://doi.org/10.1016/j.asoc.2018.12.0081

Pramokchon P, Piamsa-Nga P (2014) A feature score for classifying class-imbalanced data. In: Proceedings of the international computer science and engineering conference, https://doi.org/10.1109/ICSEC.2014.6978232

Qazi A, Goudar RH (2018) An ontology-based term weighting technique for web document categorization. Procedia Comput Sci 133:75–81. https://doi.org/10.1016/j.asoc.2018.12.0082

Qin S, Song J, Zhang P, Tan Y (2016) Feature selection for text classification based on part of speech filter and synonym merge. In: Proceedings of the international conference on fuzzy systems and knowledge discovery, https://doi.org/10.1109/FSKD.2015.7382024

Rajamohana SP, Umamaheswari K, Keerthana SV (2017) An effective hybrid cuckoo search with harmony search for review spam detection. In: Proceedings of the IEEE international conference on advances in electrical and electronics, information, communication and bio-informatics, https://doi.org/10.1109/AEEICB.2017.7972369

Rasool A, Tao R, Kamyab A (2020) GAWA - a feature selection method for hybrid sentiment classification. IEEE Access 8:191850–191861. https://doi.org/10.1016/j.asoc.2018.12.0083

Rastogi S (2018) Improving classification accuracy of automated text classifiers. In: Proceedings of the international conference on reliability, infocom technologies and optimization (Trends and Future Directions), pp 1–7

Ravi K, Ravi V (2016) Sentiment classification of Hinglish text. In: Proceedings of the international conference on recent advances in information technology, https://doi.org/10.1109/RAIT.2016.7507974

Rehman A, Javed K, Babri HA, Saeed M (2015) Relative discrimination criterion—a novel feature ranking method for text data. Expert Syst Appl. https://doi.org/10.1016/j.asoc.2018.12.0084

Rehman A, Javed K, Babri HA (2017) Feature selection based on a normalized difference measure for text classification. Inf Process Manag 53(2):473–489. https://doi.org/10.1016/j.asoc.2018.12.0085

Rehman A, Javed K, Babri HA, Asim N (2018) Selection of the most relevant terms based on a max-min ratio metric for text classification. Expert Syst Appl 114:78–96. https://doi.org/10.1016/j.asoc.2018.12.0086

Ren JS, Wang W, Wang J, Liao SS (2013) Exploring the contribution of unlabeled data in financial sentiment analysis. arXiv preprint https://doi.org/10.1016/j.asoc.2018.12.0087 pp 1149–1155

Rennie J (2019) The 20 newsgroups data set. https://doi.org/10.1016/j.asoc.2018.12.0088

Roul RK, Gugnani S, Kalpeshbhai SM (2016b) Clustering based feature selection using extreme learning machines for text classification. In: Proceedings of the IEEE international conference electronics, energy, environment, communication, computer, control, https://doi.org/10.1109/INDICON.2015.7443788

Roul RK, Bhalla A, Srivastava A (2016a) Commonality-rarity score computation. Proc Annu Meet Forum Inf Retr Eval 10(1145/3015157):3015165

Rui W, Liu J, Jia Y (2016) Unsupervised feature selection for text classification via word embedding. In: Proceedings of the IEEE international conference on big data analysis, pp 1–5, https://doi.org/10.1109/ICBDA.2016.7509787

Ruta D (2014) Robust method of sparse feature selection for multi-label classification with naive Bayes. In: Proceedings of the federated conference on computer science and information systems, pp 375–380, https://doi.org/10.15439/2014F502

Rzeniewicz J, Szymanski JS (2013) Selecting features with SVM. In: Proceedings of the iberoamerican congress on pattern recognition

Sabbah T, Selamat A, Selamat MH, Ibrahim R, Fujita H (2016) Hybridized term-weighting method for dark web classification. Neurocomputing. https://doi.org/10.1016/j.asoc.2018.12.0089

Sammut C, Webb GI (2010) Encyclopedia of machine learning. Springer, US

Sarhan AM, Hamissa GM, Elbehiry HE (2016) Proposed document frequency technique for minimizing dataset in web crawler. In: Proceedings of the international conference on computer engineering and systems, https://doi.org/10.1109/ICCES.2015.7393008

Shah FP, Patel V (2016) A review on feature selection and feature extraction for text classification. In: Proceedings of the IEEE international conference on wireless communications, signal processing and networking, https://doi.org/10.1109/WiSPNET.2016.7566545

Shahid R, Javed ST, Zafar K (2017) Feature selection based classification of sentiment analysis using biogeography optimization algorithm. In: Proceedings of the international conference on innovations in electrical engineering and computational technologies, https://doi.org/10.1109/ICIEECT.2017.7916549

Shang C, Li M, Feng S, Jiang Q, Fan J (2013) Feature selection via maximizing global information gain for text classification. Knowl Based Syst. https://doi.org/10.1016/j.knosys.2013.09.019

Shang L, Zhou Z, Liu X (2016) Particle swarm optimization-based feature selection in sentiment classification. Soft Comput. https://doi.org/10.1007/s00500-016-2093-2

Shen K, Chen X, Ke L, Lu Y, Zhang K (2013) A blended feature selection method in text. In: Proceedings of the conference on cyberspace technology, pp 573–576

Sheydaei N, Saraee M, Shahgholian A (2015) A novel feature selection method for text classification using association rules and clustering. J Inf Sci. https://doi.org/10.1177/0165551514550143

Somantri O, Kurnia DA, Sudrajat D, Rahaningsih N, Nurdiawan O, Perdana Wanti L (2019) A hybrid method based on particle swarm optimization for restaurant culinary food reviews. In: Proceedings of the international conference on informatics and computing, pp 1–5

Song Q, Ni J, Wang G (2013) A fast clustering-based feature subset selection algorithm for high-dimensional data. IEEE Trans Knowl Data Eng. https://doi.org/10.1109/TKDE.2011.181

Song J, Zhang P, Qin S, Gong J (2016) A aethod of the feature selection in hierarchical text classification based on the category discrimination and position information. In: Proceedings of the International conference on industrial informatics - computing technology, intelligent technology, industrial information integration, https://doi.org/10.1109/ICIICII.2015.116

Stambaugh C, Yang H, Breuer F (2013) Analytic feature selection for support vector machines. In: Proceedings of the machine learning and data mining in pattern recognition, pp 219–233

Sundararajan K, Palanisamy A, Versaci M (2020) Multi-rule based ensemble feature selection model for sarcasm type detection in Twitter. Comput Intell Neurosci 2020:2860479. https://doi.org/10.1155/2020/2860479

Sun J, Zhang X, Liao D, Chang V (2017) Efficient method for feature selection in text classification. In: Proceedings of international conference on engineering and technology, vol 2018-Janua, pp 1–6, https://doi.org/10.1109/ICEngTechnol.2017.8308201

Su Z, Xu H, Zhang D, Xu Y (2014) Chinese sentiment classification using a neural network tool - Word2vec. In: Proceedings of the international conference on multisensor fusion and information integration for intelligent systems, https://doi.org/10.1109/MFI.2014.6997687

Tang B, He H, Baggenstoss PM, Kay S (2016a) A Bayesian classification approach using class-specific features for text categorization. IEEE Trans Knowl Data Eng 28(6):1602–1606. https://doi.org/10.1109/TKDE.2016.2522427

Tang B, Kay S, He H (2016b) Toward optimal feature selection in naive Bayes for text categorization. IEEE Trans Knowl Data Eng. https://doi.org/10.1109/TKDE.2016.2563436

Tang B, Kay S, He H, Baggenstoss PM (2016c) EEF: exponentially embedded families with class-specific features for classification. IEEE Sig Process Lett. https://doi.org/10.1109/LSP.2016.2574327

Tang X, Dai Y, Xiang Y (2019) Feature selection based on feature interactions with application to text categorization. Expert Syst Appl 120:207–216. https://doi.org/10.1016/j.eswa.2018.11.018

Tang J, Alelyani S, Liu H (2014) Feature selection for classification: a review. Data classification: algorithms and applications

Tang B, He H (2016) FSMJ: feature selection with maximum Jensen-Shannon divergence for text categorization. In: Proceedings of the world congress on intelligent control and automation, vol 2016-Septe, pp 3143–3148, https://doi.org/10.1109/WCICA.2016.7578786

Tian W, Li J, Li H (2018) A method of feature selection based on Word2Vec in text categorization. In: Proceedings of the Chinese control conference, pp 9452–9455

Tommasel A (2016) Integrating social network structure into online feature selection. In: Proceedings of the IJCAI international joint conference on artificial intelligence, vol 2016-Janua, pp 4032–4033

Tripathy A, Anand A, Rath SK (2017) Document-level sentiment classification using hybrid machine learning approach. Knowl Inf Syst 53(3):805–831. https://doi.org/10.1016/j.knosys.2013.09.0190

Trivedi SK, Tripathi A (2017) Sentiment analyis of Indian movie review with various feature selection techniques. In: Proceedings of the IEEE international conference on advances in computer applications, https://doi.org/10.1109/ICACA.2016.7887947

Tutkan M, Ganiz MC, Akyokuş S (2016) Helmholtz principle based supervised and unsupervised feature selection methods for text mining. Inf Process Manag. https://doi.org/10.1016/j.knosys.2013.09.0191

Uysal AK (2016) An improved global feature selection scheme for text classification. Expert Syst Appl. https://doi.org/10.1016/j.knosys.2013.09.0192

Uysal AK, Gunal S (2012) A novel probabilistic feature selection method for text classification. Knowl Based Syst 36:226–235. https://doi.org/10.1016/j.knosys.2013.09.0193

Vani K, Gupta D (2017) Text plagiarism classification using syntax based linguistic features. Expert Syst Appl 88:448–464. https://doi.org/10.1016/j.knosys.2013.09.0194

Vychegzhanin SV, Razova EV, Kotelnikov EV (2019) What number of features is optimal: a new method based on approximation function for stance detection task. Proce Int Conf Inf Commun Manag ICICM 2019:43–47. https://doi.org/10.1016/j.knosys.2013.09.0195

W3Techs (2019) Historical trends in the usage of content languages for websites, September 2019. https://doi.org/10.1016/j.knosys.2013.09.0196

Wang H, Hong M (2019) Supervised Hebb rule based feature selection for text classification. Inf Process Manag 56(1):167–191. https://doi.org/10.1016/j.knosys.2013.09.0197

Wang J, Wu L, Kong J, Li Y, Zhang B (2013) Maximum weight and minimum redundancy: a novel framework for feature subset selection. Pattern Recog. https://doi.org/10.1016/j.knosys.2013.09.0198

Wang Y, Liu Y, Feng L, Zhu X (2014) Novel feature selection method based on harmony search for email classification. Knowl Based Syst. https://doi.org/10.1016/j.knosys.2013.09.0199

Wang Y, Liu Y, Zhu X (2014) Two-step based hybrid feature selection method for spam filtering. J Intell Fuzzy Syst 27:2785–2796. https://doi.org/10.1007/s00500-016-2093-20

Wang D, Zhang H, Liu R, Lv W, Wang D (2014a) T-test feature selection approach based on term frequency for text categorization. Pattern Recog Lett. https://doi.org/10.1007/s00500-016-2093-21

Wang D, Zhang H, Liu R, Liu X, Wang J (2016) Unsupervised feature selection through Gram-Schmidt orthogonalization - a word co-occurrence perspective. Neurocomputing. https://doi.org/10.1007/s00500-016-2093-22

Wang Q, Liu L, Jiang J, Jiang M, Lu Y, Pei Z (2017) Feature selection method based on multiple centrifuge models. Cluster Comput 20(2):1425–1435. https://doi.org/10.1007/s00500-016-2093-23

Wang Y, Wang J, Liao H, Chen H (2017) An efficient semi-supervised representatives feature selection algorithm based on information theory. Pattern Recog. https://doi.org/10.1007/s00500-016-2093-24

Webkb (2019) The 4 universities data set. https://doi.org/10.1007/s00500-016-2093-25

Wu L, Wang Y, Zhang S, Zhang Y (2017) Fusing gini index and term frequency for text feature selection. In: Proceedings of the IEEE international conference on multimedia big data, https://doi.org/10.1109/BigMM.2017.65

Wu G, Wang L, Zhao N, Lin H (2016) Improved expected cross entropy method for text feature selection. In: Proceedings of the international conference on computer science and mechanical automation, https://doi.org/10.1109/CSMA.2015.17

Wu G, Xu J (2016) Optimized approach of feature selection based on information gain. In: Proceedings of the international conference on computer science and mechanical automation, https://doi.org/10.1109/CSMA.2015.38

Xiaoming D, Tang Y (2013) Improved mutual information method for text feature selection. In: Proceedings of the international conference on computer science and education

Xu Z, King I, Lyu M, Jin R (2010) Discriminative semi-supervised feature selection via manifold regularization. IEEE Trans Neural Netw 21:1033–1047. https://doi.org/10.1007/s00500-016-2093-26

Xu J, Jiang H (2015) An improved information gain feature selection algorithm for SVM text classifier. In: Proceedings of the international conference on cyber-enabled distributed computing and knowledge discovery, https://doi.org/10.1109/CyberC.2015.53

Xu H, Xu L (2017) Multi-label feature selection algorithm based on label pairwise ranking comparison transformation. In: Proceedings of the international joint conference on neural networks

Yang ZT, Zheng J (2016) Research on Chinese text classification based on Word2vec. In: Proceedings of the IEEE international conference on computer and communications research

Yang J, Liu Z, Qu Z, Wang J (2014) Feature selection method based on crossed centroid for text categorization. In: Proceedings of the IEEE/ACIS international conference on software engineering, artificial intelligence, networking and parallel/distributed computing, https://doi.org/10.1109/SNPD.2014.6888675

Yang J, Lu Y, Liu Z (2019) An improved strategy of the feature selection algorithm for the text categorization. In: Proceedings of the IEEE/ACIS international conference on software engineering, artificial intelligence, networking and parallel/distributed computing, pp 3–7

Yang J, Wang J, Liu Z, Qu Z (2015) A term weighting scheme based on the measure of relevance and distinction for text categorization. In: Proceedings of the IEEE/ACIS international conference on software engineering, artificial intelligence, networking and parallel/distributed computing, https://doi.org/10.1109/SNPD.2015.7176178

Yigit F, Baykan OK (2014) A new feature selection method for text categorization based on information gain and particle swarm optimization. In: Proceedings of IEEE international conference on cloud computing and intelligence systems, https://doi.org/10.1109/CCIS.2014.7175792

Yousefpour A, Ibrahim R, Hamed HNA (2017) Ordinal-based and frequency-based integration of feature selection methods for sentiment analysis. Expert Syst Appl 75:80–93. https://doi.org/10.1007/s00500-016-2093-27

Zainuddin N, Selamat A, Ibrahim R (2018) Hybrid sentiment classification on Twitter aspect-based sentiment analysis. Appl Intell 48(5):1218–1232. https://doi.org/10.1007/s00500-016-2093-28

Zhang Z, Ke T, Deng N, Tan J (2014) Biased p-norm support vector machine for PU learning. Neurocomputing 136:256–261. https://doi.org/10.1007/s00500-016-2093-29

Zhang J, Hu X, Li P, He W, Zhang Y, Li H (2014a) A hybrid feature selection approach by correlation-based filters and SVM-RFE. In: Proceedings of the international conference on pattern recognition, pp 3684–3689, https://doi.org/10.1109/ICPR.2014.633

Zhang H, Ren YG, Yang X (2013) Research on text feature selection algorithm based on information gain and feature relation tree. In: Proceedings of the web information system and application conference, pp 446–449, https://doi.org/10.1109/WISA.2013.90

Zhen Z, Wang H, Xing Y, Han L (2016) Text feature selection approach by means of class difference. In: Proceedings of the international conference on natural computation, fuzzy systems and knowledge discovery, https://doi.org/10.1109/FSKD.2016.7603412

Zhou X, Hu Y, Guo L (2014) Text categorization based on clustering feature selection. Procedia Comput Sci 31:398–405. https://doi.org/10.1177/016555151455014330

Zhou H, Han S, Liu Y (2018) A novel feature selection approach based on document frequency of segmented term frequency. IEEE Access 6:53811–53821

Zhou H, Guo J, Wang Y, Zhao M (2016) A feature selection approach based on interclass and intraclass relative contributions of terms. Comput Intell Neurosci. https://doi.org/10.1155/2016/1715780

Zhu L, Wang G, Zou X (2017) Improved information gain feature selection method for Chinese text classification based on word embedding. Proc Int Conf Softw Comput Appl 10(1145/3056662):3056671

Zhuang Y, Wang H, Xiao J, Wu F, Yang Y, Lu W, Zhang Z (2017) Bag-of-discriminative-words (BoDW) representation via topic modeling. IEEE Trans Knowl Data Eng 29(5):977–990. https://doi.org/10.1109/TKDE.2017.2658571

Zong W, Wu F, Chu LK, Sculli D (2015) A discriminative and semantic feature selection method for text categorization. Int J Prod Econ 165:215–222. https://doi.org/10.1016/j.ijpe.2014.12.035

Zuo Z, Li J, Anderson P, Yang L, Naik N (2018) Grooming detection using fuzzy-rough feature selection and text classification. In: Proceedings of the IEEE international conference on fuzzy systems, pp 1–8