Expert Systems With Applications 106 (2018) 36–54

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

A recent overview of the state-of-the-art elements of text classification

Marcin Michał Mirończuk*, Jarosław Protasiewicz

National Information Processing Institute, al. Niepodległości 188 b, 00-608 Warsaw, Poland

ARTICLE INFO

Article history:
Received 23 November 2017
Revised 26 March 2018
Accepted 27 March 2018
Available online 28 March 2018

Keywords:
Text classification
Document classification
Text classification overview
Document classification overview

ABSTRACT

The aim of this study is to provide an overview the state-of-the-art elements of text classification. For this purpose, we first select and investigate the primary and recent studies and objectives in this field. Next, we examine the state-of-the-art elements of text classification. In the following steps, we qualitatively and quantitatively analyse the related works. Herein, we describe six baseline elements of text classification including data collection, data analysis for labelling, feature construction and weighing, feature selection and projection, training of a classification model, and solution evaluation. This study will help readers acquire the necessary information about these elements and their associated techniques. Thus, we believe that this study will assist other researchers and professionals to propose new studies in the field of text classification.

1. Introduction

Text classification is a construction problem of models which can classify new documents into pre-defined classes (Liu, 2006; Manning, Raghavan, & Schütze, 2008). Currently, it 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. Text classification remains a prominent research topic, utilising various techniques and their combinations in complex systems. Furthermore, researchers are either developing new classification systems or improving the existing ones, including their elements to yield better results, i.e. a higher computational efficiency (Altınel, Ganiz, & Diri, 2015; Pinheiro, Cavalcanti, & Tsang, 2017; Wang, Wu, Zhang, Xu, & Lin, 2013).

Literature overviews of text classification usually reveal its crucial elements, techniques, and solutions, proposing the further development of this research area. Nevertheless, the existing reviews are still useful as they address the significant problems of text classification (Aas & Eikvil, 1999; Aggarwal & Zhai, 2012). However, these works are slightly outdated as they do not include the latest studies. Furthermore, their explanation of text classification has some limitations, for example, they lay emphasis only on machine learning techniques or algorithms, omit some essential elements of text classification, or focus on a particular research domain (Adeva, Atxa, Carrillo, & Zengotitabengoa, 2014; Guzella & Caminhas, 2009; Ittoo, Nguyen, & van den Bosch, 2016). We reiterate here that these are excellent works, which are still useful to the research community. However, with the increasing interest in the area of text classification, we need the most recent systematic overview to better understand what has been achieved in this field.

In this study, we aim to overcome the difficulties mentioned above. Moreover, the article presents a latest and holistic summary of text classification. We direct significant effort to generate a research map for text classification to assist in recognising its main elements by examining both the most recent and former studies. More specifically, in addition to the understandable requirement to complement the existing reviews, the objectives of this study are as follows:

1. To extract and present the essential phases of the text classification process, including the most common vocabulary, as a baseline framework. This framework could be referred as the map of text classification.
2. To enumerate both the older and new techniques utilised in each phase of text classification. These techniques are identified systematically via a qualitative analysis.
3. To perform a quantitative analysis of the system to exhibit the research trends in this area.

According to the best of our knowledge, there are no similar recent studies in the form of an overview of the investigated field. Furthermore, we believe that this study significantly systematises and enhances the knowledge regarding the modelling of classifi-

* Corresponding author.
E-mail addresses: marcin.mironczuk@opi.org.pl (M.M. Mirończuk), jaroslaw.protasiewicz@opi.org.pl (J. Protasiewicz).

https://doi.org/10.1016/j.eswa.2018.03.058

Table 1
Text classification elements grouped based on the category of a text classification issue and arranged in ascending order of years. Label 'Yes' implies that the article mentioned or introduced the corresponding text classification element; label 'No' implies the opposite case.

Text classification element			Article			
–			Aas and Eikvil (1999)	Aggarwal and Zhai (2012)	Guzella and Caminhas (2009)	Adeva et al. (2014)
Pre-processing			Yes	Yes	Yes	Yes
Vector space model			Yes	Yes	Yes	Yes
Dimensionality reduction	Feature selection	Document frequency thresholding	Yes	Yes	Yes	Yes
		Information gain	Yes	Yes	Yes	No
		χ^2 statistic	Yes	Yes	Yes	Yes
		Gini index	No	Yes	No	No
		Mutual information	No	Yes	No	No
		Odds ratio	No	No	Yes	No
		Term-frequency variance	No	No	Yes	No
	Feature projection	Singular value decomposition	Yes	No	No	No
		Different type of LSI	No	Yes	No	No
		Supervised clustering for dimensionality reduction	No	Yes	No	No
		Linear discriminant analysis	No	Yes	No	No
		Generalised singular value decomposition	No	Yes	No	No
Training method of a classification function	Rocchio's algorithm		Yes	Yes	No	Yes
	Naive Bayes		Yes	Yes	Yes	Yes
	k-nearest neighbour		Yes	Yes	Yes	Yes
	Decision trees		Yes	Yes	Yes	No
	Support vector machines		Yes	Yes	Yes	Yes
	Rule-based classifiers		No	Yes	No	No
	Regression-based classifiers		No	Yes	Yes	No
	Neural network classifiers		No	Yes	Yes	No
	The ensemble learning techniques		Yes	Yes	Yes	No
	Artificial immune systems		No	No	Yes	No
Performance measures/Evaluation measurement technique			Yes	Yes	Yes	Yes
Dataset(s) (mentioned or introduced)			Yes	Yes	Yes	Yes
Description of domain-specific difficulties			No	No	Yes	Yes

cation systems. The results of the text classification process with its elements are particularly relevant. Moreover, we show that it is possible to identify, explore, and develop new aspects of text classification or alternatively upgrade its existing components. In addition, our study constitutes a relevant and modern complement to the current reviews.

The paper is structured as follows. Section 2 presents a comprehensive description of the existing reviews. Next, Section 3 describes the text classification process and explains the review procedure. Then, Section 4 explains the problems, objective, and components of text classification via a qualitative analysis. Section 5 introduces a quantitative analysis of the text classification journals, including conference proceedings. Finally, Section 6 concludes the research study.

2. Related works

The accessible reviews mostly describe and focus on the following five elements of the text classification process: (1) document pre-processing, i.e. tokenisation, stop-word removal, and stemming or lemmatisation, (2) document modelling, i.e. representing a document in an appropriate form, to be processed by a machine learning algorithm, (3) feature selection and projection, (4) machine learning algorithm utilisation to construct a classification model or function, and (5) quality indicators and evaluation methods. Table 1 lists all these elements and the review works directly related to them. In addition, the table notes dataset issues and domain-specific difficulties.

Aas and Eikvil (1999) have previously enumerated and described the text classification steps, namely, pre-processing, vector space model creation, dimensionality reduction (feature selection and projection), training of a classification function, and performance measurement. In their work, text classification was used to present several schemas of feature weighting, e.g. Boolean, term frequency, inverse document frequency, and entropy. Moreover, the authors explained three feature selection methods, namely, document frequency thresholding, information gain, and χ^2-statistic, and one feature projection method, namely, latent semantic indexing (LSI). In addition, they summarised and elucidated six machine learning methods: Rocchio's algorithm, naive Bayes, k-nearest neighbour, decision tree, support vector machine (SVM), and ensemble learning, including bagging and boosting algorithms. Furthermore, they described the performance measures for binary, multi-class, and multi-label classification tasks. They also provided a Reuters-21578 dataset, suitable for different classification experiments.

Aggarwal and Zhai (2012) described text processing similarly to Aas and Eikvil (1999), but they provided more examples. Additionally, the authors presented a more in depth discussion on each element of the text classification process. Their paper starts with a description of the classification problems, software, and examples of the domains in which text classification is commonly used. In the article, first, the authors introduce (1) feature selection methods, including the Gini index, information gain, mutual information, and χ^2-statistic, and (2) feature projection methods, such as different types of LSI, supervised clustering for dimension-

ality reduction, linear discriminant analysis, and generalised singular value decomposition. Second, they describe the following different types of classification learning algorithms: decision tree, rule-based classifiers, naive Bayes (multivariate Bernoulli and multinomial models), SVM, regression-based classifiers, neural network, and proximity-based classifiers (k-nearest neighbours, Rocchio's). In addition, they discuss ensemble learning techniques, including simple committees, boosting, and bagging. Finally, Aggarwal and Zhai (2012) explain the measures of accuracy of the classification process. We also notice some implicit and explicit observations made in their paper regarding the classification techniques. These observations are related to, for example, different types of classification tasks and their solutions and performance of the linear classifiers. Moreover, it is worth mentioning that Aggarwal and Zhai (2012) address the interesting problems of linked and web data classification.

The last two works of Guzella and Caminhas (2009) and Adeva et al. (2014) in Table 1 present a classification process similar to the studies mentioned above. However, the authors focus on solutions for domain-specific problems. Guzella and Caminhas (2009) widely discussed the issue of spam filtering. In their work, we notice an extended description of text processing elements, such as datasets, document representation (e.g. n-grams techniques), performance measurement, and comparative studies of spam filtering methods. Moreover, the authors state and describe the problem of a concept drift, which other researchers have not considered. The paper by Adeva et al. (2014) explains the approach for improving systematic medical reviews. For this purpose, the authors briefly describe all the elements of the classification system and present a text corpus for the experiments for the evaluation of the proposed system.

All the works discussed above contribute significantly to text classification. However, the description of the text classification process needs further improvement. We assumed that we can deliver a more comprehensive, holistic, and organised schema, including a dictionary, to deliberate on the text classification problems. Moreover, in our view, there is a requirement for a broader but not necessarily a deeper perspective of the text classification process which focuses on appropriately grouped domain keywords. Based on this perspective, we may better understand the level of advancement which has been attained in the field of text classification.

Moreover, the text classification process herein is more sophisticated and complex than the previously discussed examples. It includes six primary elements: (1) data acquisition, (2) data analysis and labelling, (3) feature construction and weighting, (4) feature selection and/or projection, (5) model training, and (6) solution evaluation. All these elements should be better justified. Furthermore, the techniques related to each component should be classified appropriately.

Finally, it is worth mentioning that a learning method may utilise different learning algorithms to train a classification model or function used to classify documents into defined classes. Numerous well-described and well-known machine learning algorithms are suitable for this purpose, e.g. artificial neural network, k-nearest neighbour method, decision tree, decision rules (rule-based classifiers), naive Bayes, cost-sensitive selective naive Bayes, and support vector machines, etc. (Aggarwal, 2015; Aggarwal & Zhai, 2012; Ibáñez, Bielza, & Larrañaga, 2014; Nunzio, 2014; Witten, Frank, Hall, & Pal, 2016). Therefore, herein, a discussion on this issue is not presentence; instead, we focus on more general learning approaches which, in our view, have not been well-justified in the reviews mentioned above.

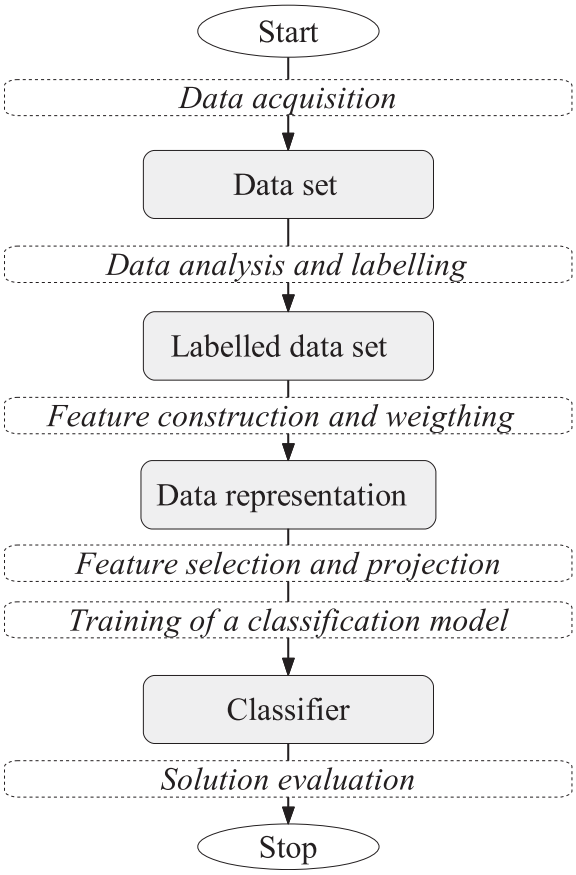

Fig. 1. Flowchart of the text classification process with the state-of-the-art elements.

3. Overview of a review method

In this section, we present a brief outline of the baseline framework for text classification along with our review method for the studies.

3.1. Text classification framework

The investigated baseline process for text classification includes the six elements mentioned in Section 2. Fig. 1 presents a flowchart of this process, which we discuss briefly below.

As we can see in Fig. 1, the classification process starts with data acquisition from various text sources, including internal datasets, the Internet, and open databases. From the data acquisition, we obtain a dataset representing a physical or business process. Next, the dataset is pre-processed to generate a representation required by the selected learning method. This non-trivial issue consists of two phases. First, the features are constructed from the data. Then, they are accordingly weighted with the selected feature representation algorithm to yield the appropriate data representation. Then, the number of features are reduced by the feature selection method. Subsequently, the reduced features are projected onto a lower dimensionality to achieve the optimal data representation. Following this, different learning approaches are used to train a classification function to recognise a target concept. When a classification model is adequately developed, it can classify incoming data that have to be represented in a manner similar to in the training phase. Consequently, the classifier produces a decision that defines the class of each input vector. Technically, the decisions are probabilities or weights. Finally, the evaluation procedure is utilised to estimate the text classification pro-

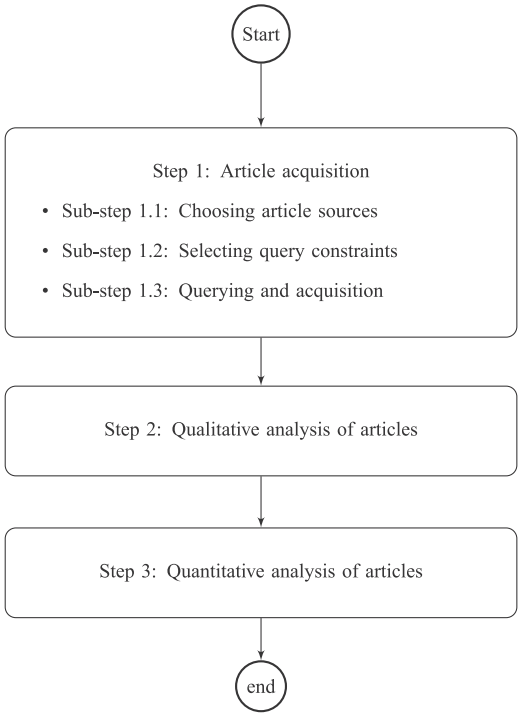

Fig. 2. Graphical representation of the review method for the studies.

cess operation. All the above phases are the main elements of the framework for text classification.

3.2. Review method

We apply a review method which included three phases, namely, (1) article acquisition, (2) analysis of the quality of the articles, i.e. assigning them to the appropriate steps of the text classification process, and (3) quantitative analysis of the articles. Fig. 2 presents the above phases and their sub-steps.

It can be seen from the figure that the phase of article acquisition is composed of three sub-steps. First, the article sources are selected. In practice, we utilised Springer and Elsevier sources most extensively. In addition, we used ACM and IEEE repositories but much less broadly. We have to mention that we excluded arXiv which is undoubtedly a well-known repository, but mostly it includes only pre-print manuscript versions. They are merely moderated instead of being reviewed. We should be aware that some authors submit their works to arXiv and a peer-reviewed journal concurrently. If a particular manuscript fails a review pipeline, arXiv includes only the unverified study containing questionable results. Since the results of our study are the guidelines in text classification, we based research only on repositories, which contains thoroughly reviewed and well-established journals. For these reasons, arXiv could not be included in the literature review. We made an exception from this rule only for one literature review paper. Second, the search constraints regarding the article type and publication year are set. We focused only on regular papers published in journals, rather than conference articles. Occasionally, we included books and conference papers if there was the lack of relevant articles from journals or they were strongly related to text classification. The baseline range of the publication year was between 2013 and 2018. A lower range was used if a query did not return any relevant articles. In the final sub-step, for each primary element of text classification, several keywords which describe them most accurately are selected. Keywords with the constraints are included in queries that are sent to search engines, and accordingly, their responses are acquired. The achieved articles are stored for manual review including analysis. Table 2 presents the request examples for each investigated category and our preliminary comments regarding the returned results.

In the phase of qualitative analysis of the articles, we manually reviewed the collected papers. In addition, we created a taxonomy in this step. Each article was assigned to the appropriate element of the text classification framework. Thus, we generated a database of classified articles. This database is freely available as an open dataset repository.[1] In some cases, an article might be related to more than one category.

The phase of quantitative analysis of articles included a quantitative study of the research on text classification. It was based on the collected manuscripts and their taxonomy. Thus, we could identify the elements of the classification framework which were explored most thoroughly.

Table 2 also lists the request examples for baseline keywords. We used more general terms when the requests could not find articles. Therefore, we could highlight research topics that were at no time extensively investigated in the articles on text classification. Thus, we consider the above keywords as the basic dictionary or vocabulary of text classification.

4. Qualitative analysis of text classification studies

This section extensively outlines the literature related to text classification. More specifically, we explain the problem and objectives of text classification. Furthermore, we discuss the main components of the text classification process, learning methods, and evaluation approaches. The conducted analysis establishes the methods which are used in text classification and identifies the still unexplored areas of the application.

4.1. Problem and objectives of text classification

Text document classification (text classification) is the problem of assigning predefined classes (labels) to an unlabelled text document. Numerous studies present different approaches and applications of text classification. The various categories of text classification are domain, classification purpose, classification task, general approach, and dedicated approach. Table 3 presents some details about these categories.

First, we found that several works, including surveys, applied the classification task only in a particular domain, where the domain is considered as a source of the textual data for the classification. Table 3 lists several domains, such as industry, finance, and medicine. Second, the studies which were determined to have an explicit reason for text classification were placed in the classification purpose category. For example, the reasons may be recognition of commercial or non-commercial websites and academic or researcher home pages. Third, the classification task category includes articles related to the topics that consider binary, multi-class, multi-label, and hierarchical classification problems. Briefly, a binary classifiers model a two-class problem and classify objects into one class. A multi-class classification problem models a classification problem with more than one class. A binary classification problem is a multi-class classification problem with two 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. The hierarchical classification problem may be seen as a multi-class classification problem (each node is a class) and a multi-label classification problem (each path from the root to the leaf is a set of labels for the object

[1] https://doi.org/10.5281/zenodo.1207374

Table 2
Request examples sent to the search engines of publishers. We have acquired 242 articles; however, the table contains 233 papers, since we excluded nine studies that are literature reviews or unclassified materials.

Category request	Example request	Comment	No. of articles
Classification systems and application area	Sentiment analysis; website classification; spam identification	The requests return numerous works. It appears that this is a prominent research area.	60
Labelling methods	Bags of words text classification; bags of feature text classification; text labelling methods	The requests return articles related to multiple instance learning.	8
Feature construction	Text feature construction methods; feature construction in text; feature construction review; construct features from text; feature construction in text mining; information extraction text classification; ontology for text classification; taxonomies in text classification; taxonomy for text classification; WordNet for text classification; keywords based text classification	The requests return only a few works related to feature construction. It appears that this topic is largely not explained in the articles.	39
Feature weighting	Text feature weight	The requests return numerous works. It appears that this is a prominent research area.	14
Feature selection	Review feature selection; feature selection; text classification feature selection	There are various studies dealing with feature selection. In addition, there are some general review articles, but they are not directly related to text classification.	38
Feature projection	Feature projection; text classification pca; text classification principal component analysis; text classification dimension reduction; review feature projection; review dimension reduction	There are studies tackling with future projection, but they are fewer than those on feature selection. Additionally, there are some review articles, but they are not directly related to text classification.	19
Classification method/Learning methods	Text classification learning methods; supervised text classification; semi supervised text classification; ensemble learning text classification; the method of combine text classification functions; multiview text classification	There are numerous papers on the issue of training algorithms in text classification.	43
Solution evaluation	text classification evaluation; text classification indicators; evaluation methods	There are various articles related to the problem of solution evaluation in text classification.	12

Table 3
Various categories of text classification.

Category name	Topics and selected works
Domain	- Industry (Ittoo et al., 2016; Lin, 2009), - Finance (de Fortuny, Smedt, Martens, & Daelemans, 2014; Kumar & Ravi, 2016; Li, Xie, Chen, Wang, & Deng, 2014), - Medicine (Adeva et al., 2014; Mostafa & Lam, 2000; Parlak & Uysal, 2015; Shen et al., 2016), - Internet, such as analysis of email, server logs, web pages, websites, and tweets. (Basto-Fernandes et al., 2016; Chang & Poon, 2009; Cuzzola, Jovanović, Bagheri, & Gašević, 2015; Kan & Thi, 2005; Qi & Davison, 2009), - Patent databases (Giachanou, Salampasis, & Paltoglou, 2015; Li & Shawe-Taylor, 2007; Trappey, Hsu, Trappey, & Lin, 2006).
Classification purpose	- Classification of emotions, sentiments, and opinions (Abbasi, Chen, & Salem, 2008; Catal & Nangir, 2017; García-Pablos et al., 2018; Giatsoglou et al., 2017; Kang et al., 2018; Li et al., 2017a; Onan, Korukoğlu, & Bulut, 2016b; Perikos & Hatzilygeroudis, 2016; Vinodhini & Chandrasekaran, 2016; Xia, Xu, Yu, Qi, & Cambria, 2016; Xia, Zong, & Li, 2011), - Identification of academic or researcher home pages (Gollapalli, Caragea, Mitra, & Giles, 2013; 2015; Gollapalli, Giles, Mitra, & Caragea, 2011), - Support a selection of proper web pages in focused crawling (Saleh, Abulwafa, & Al Rahmawy, 2017a), - Tweet language identification (Castro, Souza, Vitório, Santos, & Oliveira, 2017), - Support automation of systematic reviews (Karystianis, Thayer, Wolfe, & Tsafnat, 2017), - Categorisation parts of PDF text (Bui, Del Fiol, & Jonnalagadda, 2016), - Detection of irony (Charalampakis, Spathis, Kouslis, & Kermanidis, 2016), - Detection of terrorism-related articles on the web (Choi, Ko, Kim, & Kim, 2014), - Categorisation of applied science research projects according to the corresponding technologies of research funding organisations (Thorleuchter & den Poel, 2013), - Recognition of commercial or non-commercial websites (Qi & Davison, 2009).
Classification task	- Binary, multi-class, multi-label, and hierarchical (Al-Salemi, Noah, & Ab Aziz, 2016; Altınçay & Erenel, 2014; Du, Liu, Ke, & Gong, 2018; Elghazel et al., 2016; Gui et al., 2012; Jiang, Li, & Pan, 2016b; Kang, Cho, & Kang, 2015; Pillai, Fumera, & Roli, 2017; Ruiz & Srinivasan, 2002; Sokolova & Lapalme, 2009)
General approach	- Tuning of naive Bayesian classifiers (Diab & El Hindi, 2017), - Creation of a lightweight text classifier (Silva, Almeida, & Yamakami, 2017), - Creation of an interpretable classification method Van Linh, Anh, Than, and Dang (2017), - Building a scalable and an effective text classifier (Pang, Jin, & Jiang, 2015), - Development of novel prototype-based classifiers (Pang & Jiang, 2013; Zhang, Chen, & Guo, 2013), - Improvement of the implementation of associative classifiers for document classification (Yoon & Lee, 2007).
Dedicated approach	- Investigation of mobile text classification (Yin & Xi, 2017), - Investigation of Hadiths (textual sources of law) (Saloot et al., 2016), - Study of SMS spam filtering (Almeida, Silva, Santos, & Hidalgo, 2016), - Inspection of Dark Web classification (Sabbah, Selamat, Selamat, Ibrahim, & Fujita, 2016).

of that node). Fourth, we added the works focused on the general approaches to text classification to the class called the general approach. The group includes for example studies on tuning naive Bayesian classifiers and creation of a lightweight text classifier. Finally, in the last category, dedicated approach, we placed some papers dealing with particular or dedicated text classification problems. In the last group, we have studies resolving the text classification problem based on well-defined text data sources such as an SMS dataset for spam filtering and Hadiths.

4.2. Qualitative analysis of studies

In this subsection, we discuss the elements of text classification based on the framework depicted in Fig. 1.

4.2.1. Data acquisition

The data acquisition stage is executed when we do not already have the necessary data. In this phase, we acquire data required to solve a stated research objective, i.e. an assumed research hypothesis related to a classification task. There are several open data sets, e.g. Reuters, TDT2, WebKB, and Newsgroup (Elghazel, Aussem, Gharroudi, & Saadaoui, 2016; Lochter, Zanetti, Reller, & Almeida, 2016; Pinheiro et al., 2017; Pinheiro, Cavalcanti, & Ren, 2015; Sabbah et al., 2017). Pinheiro et al. (2017) have enumerated as well as used in their research 47 different datasets. We have to underline that datasets related to solving particular classification purpose are publicly available. For example, datasets from STS-Test, HCR, and IPhone6 were created to detect views in short text messages (Lochter et al., 2016).

4.2.2. Data analysis and labelling

During the phase of data analysis and labelling, mainly labelled data-sets are prepared. There are two strategies for data labelling: labelling groups of texts and assigning a label or labels to each text part. The first strategy is called multi-instance learning (Alpaydın, Cheplygina, Loog, & Tax, 2015; Foulds & Frank, 2010; Herrera et al., 2016; Liu, Xiao, & Hao, 2018; Ray, Scott, & Blockeel, 2010; Xiao, Liu, Yin, & Hao, 2017; Yan, Li, & Zhang, 2016; Yan, Zhu, Liu, & Wu, 2017), whereas the second one includes different supervised methods (Sammut & Webb, 2017). The results of this phase are employed in the succeeding stages.

4.2.3. Feature construction and weighting

In this phase, the labelled data set is represented appropriately for a learning algorithm. Fig. 3 shows the basic steps of this phase. Two well-known representations of textual data are:

- vector space representation or model (Hand, Mannila, & Smyth, 2001; Manning et al., 2008), where a document is represented as a vector of feature weights, whereas the words (terms, phrases) in the document form features,
- graph representation (Mihalcea & Radev, 2011; Schenker, Bunke, Last, & Kandel, 2005), where a document is modelled in a graph form, for example, nodes represent words, whereas edges represent the relationships between the words.

Both representations are based on features and their weights. In the literature, there are numerous approaches for their generation. The most distinctive feature types are as follows:

- simple features (keywords or phrases), including uni-grams, bi-grams, and n-grams (Chang & Poon, 2009; Figueiredo et al., 2011; Lee, Isa, Choo, & Chue, 2012; Onan, Korukoglu, & Bulut, 2016a; Xie, Wu, & Zhu, 2017),
- taxonomies or ontologies of features (Cagliero & Garza, 2013; Kang, Haghighi, & Burstein, 2016; de Knijff, Frasincar, & Hogenboom, 2013; Li, Yang, & Park, 2012; Liu, He, Lim, & Wang, 2014; Pi, Martí, & Garcia, 2016; Saleh, Al Rahmawy, & Abulwafa, 2017b; Wu et al., 2017),
- domain specific features, for example opinion, emotion or aspect lexicons/words (Agathangelou, Katakis, Koutoulakis, Kokkoras, & Gunopulos, 2018; Bandhakavi, Wiratunga, Padmanabhan, & Massie, 2017; Manek, Shenoy, Mohan, & Venugopal, 2017),
- embedded features (Baroni, Dinu, & Kruszewski, 2014), i.e. words to vectors (Word2vec) (Chaturvedi, Ong, Tsang, Welsch, & Cambria, 2016; Enríquez, Troyano, & López-Solaz, 2016; Mikolov, Yih, & Zweig, 2013; Tommasel & Godoy, 2018; Wang et al., 2016) or global vectors for word representation (GloVe) (Pennington, Socher, & Manning, 2014),
- simple semantic features, e.g. named entities and noun phrases (Gui, Gao, Li, & Yang, 2012; Král, 2014; Li et al., 2012; Saha & Ekbal, 2013),
- features extracted based on topic modelling (Pavlinek & Podgorelec, 2017; Qin, Cong, & Wan, 2016; Zhang & Zhong, 2016; Zuo, Zhao, & Xu, 2016), or dissimilarity space (Duin, Loog, Pękalska, & Tax, 2010; Pinheiro et al., 2017),
- other meta information, such as Wikipedia knowledge (Wang, Hu, Zeng, & Chen, 2009) or sentence importance (Ko, Park, & Seo, 2004).

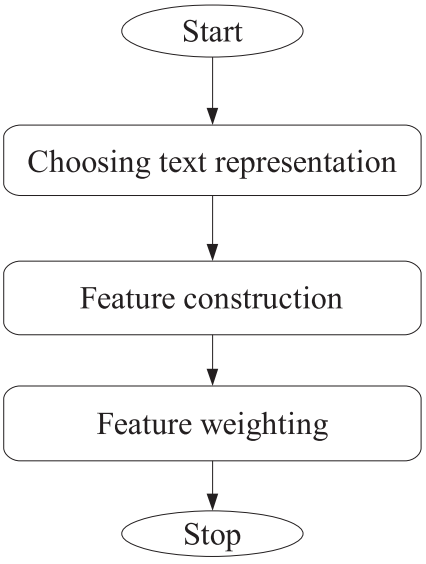

Fig. 3. Basic steps of text document transformation.

Only after the features are selected, numerical values can be assigned to them. The feature weighting issue including its impact on text classification is widely discussed in the literature. Currently, there are well-known and widely used schemes of feature weighting in the text processing and classification fields. Moreover, there are some recent and probably less known feature weighting methods. Table 4 lists both the older and latest schemes of feature weighting.

4.2.4. Dimensionality reduction

The dimensionality reduction phase is performed when numerical values are assigned to features. This step can be seen as a sort of data compression, which mainly is realised in two steps. First, the feature selection methods select the most important features. Next, the feature project techniques transform an existing feature space into another one (Vlachos, 2010). Fig. 4 shows the different types of dimensionality reduction.

Table 4
Known and used schemes of feature weighting.

Time category	Schema name of feature weighting
Older and well-known feature weighting methods	- Binary, - Term frequency (*tf*), - Inverse document frequency (*idf*), - Term frequency-inverse document frequency $tf \cdot idf$, - The BM25 weighting scheme is a state-of-the-art term weighting and retrieval model, often known as Okapi weighting or Okapi BM25 (BM stands for Best Matching and 25 is the scheme version number). The schemes mentioned above were well described by Haddoud, Mokhtari, Lecroq, and Abdeddaïm (2016), Manning et al. (2008), Robertson, Walker, Jones, Hancock-Beaulieu, and Gatford (1995).
Recent and less known feature weighting methods	- LGT scheme (Rao et al., 2017), - Modified frequency-based term weighting schemes for text classification (Sabbah et al., 2017), - Novel adaptive feature weighting approaches for naive Bayes text classifiers (Zhang, Jiang, Li, & Kong, 2016), - Term frequency and inverse gravity moment (Chen, Zhang, Long, & Zhang, 2016), - Model-induced term-weighting schemes (Kim & Kim, 2016), - Deep feature weighting (Jiang, Li, Wang, & Zhang, 2016a), - Term weighting based on class density (Fattah, 2015), - LI binary and LI frequency (Ke, 2015), - Genetic algorithms to combine a set of basic weights to generate discriminative term-weighting schemes (Escalante et al., 2015), - Class-indexing-based term weighting (Ren & Sohrab, 2013), - Semantic term weighting (Luo, Chen, & Xiong, 2011).

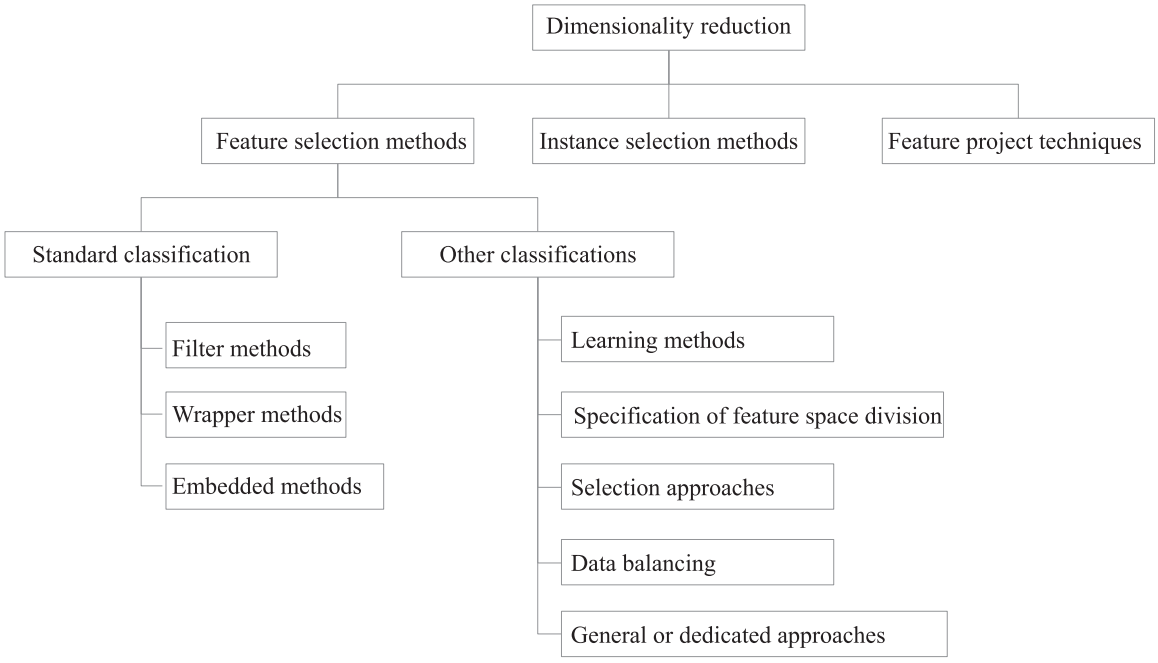

Fig. 4. Different types of dimensionality reduction.

Feature selection. The feature selection techniques retain only the most relevant/descriptive features or dimensions and discard the remaining. The standard review approach classifies them into three main groups, namely (Bellotti, Nouretdinov, Yang, & Gammerman, 2014; Guyon & Elisseeff, 2003; Li, Li, & Liu, 2017b) filter, wrapper, or embedded feature selection methods. The filter methods utilise a feature ranking function to select the best features. The ranking function assigns a relevance score based on a sequence of examples. Intuitively, a more relevant feature will be higher in rank. Subsequently, the *n*-top features are retained or the *n*-worst features are removed from the dataset. The wrapper methods are general-purpose algorithms employed for searching the space of feature subsets. Furthermore, the wrapper methods test the performance of each subset using a learning algorithm. Finally, the feature subset that yields the best performance is selected for use. The embedded feature selection methods learn which features contribute most to the accuracy of a model while the model is being created. Some learning algorithms (e.g. decision trees) include an embedded feature selection method so that it becomes an implicit part of their learning process.

Some of the well-known state-of-the-art implementations of the above techniques are Fisher ranking, correlation coefficients, mutual information, normalised punctual mutual information, χ^2, Kolmogorov–Smirnov test, Mann–Whitney *U* test, LASSO, elastic net, and ridge regression (Bouma, 2009; Li, Xia, Zong, & Huang, 2009; Manning et al., 2008; Schürmann, 1996; Vergara & Estévez, 2014; Zhou, Jin, & Hoi, 2010). Also, we can enumerate solutions that were projected in special for text classification, like:

- multivariate relative discrimination criterion (MRDC) (Labani, Moradi, Ahmadizar, & Jalili, 2018),
- feature unionisation (Jalilvand & Salim, 2017),
- searching for discriminative words in multidimensional continuous feature space (Sajgalik, Barla, & Bielikova, 2017),
- normalised difference measure (NDM) (Rehman, Javed, & Babri, 2017),
- variable global feature selection scheme (VGFSS) (Agnihotri, Verma, & Tripathi, 2017),
- meaning based feature selection (MBFS) (Tutkan, Ganiz, & Akyokuş, 2016),
- hybrid feature selection based on enhanced genetic algorithm (Ghareb, Bakar, & Hamdan, 2016),
- improved global feature selection scheme (IGFSS) (Uysal, 2016).

In addition to the above approaches, we can examine the following other types of feature selection:

- learning methods, e.g. semi-supervised approaches (Sheikhpour, Sarram, Gharaghani, & Chahooki, 2017; Wang, Wang, Liao, & Chen, 2017) and ensemble techniques (Ravi & Ravi, 2017; Seijo-Pardo, Porto-Díaz, Bolón-Canedo, & Alonso-Betanzos, 2017),
- specification of feature space division, e.g., global, local, or class-specific feature selection approaches (Armanfard, Reilly, & Komeili, 2016; Pinheiro, Cavalcanti, Correa, & Ren, 2012; Pinheiro et al., 2015; Tang, He, Baggenstoss, & Kay, 2016a; Tang, Kay, & He, 2016b),
- classification tasks, e.g. binary-class (Badawi & Altinçay, 2014), multi-class (Tang et al., 2016b), or multi-label (Alalga, Benabdeslem, & Taleb, 2016; Pereira, Plastino, Zadrozny, & Merschmann, 2018),
- data balancing (Ogura, Amano, & Kondo, 2011),
- general and dedicated approaches (Liu, Bi, & Fan, 2017; Liu, Wang, Feng, & Zhu, 2016; Meng, Lin, & Yu, 2011; Parmezan, Lee, & Wu, 2017; Uysal & Günal, 2012; Wang, Li, Song, Wei, & Li, 2011; Yu, Wu, Ding, & Pei, 2016).

The research field of feature selection is extensively explored by researchers, and elaborated methods are widely applied in practice. Instance selection methods are also known apart from feature selection. In this approach, the space of instances rather than the space of features is reduced (Olvera-López, Carrasco-Ochoa, Trinidad, & Kittler, 2010; Tsai & Chang, 2013).

Feature projection. Feature projection methodologies project the existing features onto different dimensions. The aim here is to obtain new data axes so that the new dataset structure and its variance retain the original dataset structure as closely as possible (Burges, 2010; Cunningham & Ghahramani, 2015). Herein, we list several multidimensional scaling techniques, including (1) Convex sparse PCA (CSPCA) (Chang, Nie, Yang, Zhang, & Huang, 2016), (2) Imprecise spectrum analysis (ISA) for a linear spectral analysis of documents (Guan, Zhou, Xiao, Guo, & Yang, 2013), (3) t-Distributed stochastic neighbour embedding (t-SNE) (van der Maaten, 2014; van der Maaten & Hinton, 2008), (4) LSI (Kontostathis & Pottenger, 2006), which is based on singular value decomposition (SVD), (5) Kernel principal component analysis (nonlinear PCA) (Schölkopf, Smola, & Müller, 1998), (6) Linear discriminant analysis (LDA) (Ripley & Hjort, 1995), (7) Linear principal component analysis (PCA) (Mardia, Kent, & Bibby, 1979), and (8) Sammon (1969).

It is worth mentioning that feature projection uses cases that are extensively discussed in the literature, for example,

- email classification (Gomez & Moens, 2012),
- research on new methods for text classification (Chen, Guo, & Wang, 2011),
- improvement of feature selection methods (Uguz, 2011),
- similarity detection between patent documents and scientific publications (Magerman, Looy, & Song, 2010),
- general studies on the impact of feature projection on text classification (Kim, Howland, & Park, 2005; Li & Park, 2009; Wang & Yu, 2009).

4.2.5. Training of classification models

Several learning approaches are used to train a classification function to recognise a target concept. This stage is executed only if an appropriately prepared training set is available. Training algorithms can be grouped into supervised, semi-supervised, ensemble, active, transfer, or multi-view learning approaches.

Supervised learning. It refers to any machine learning process that trains a function by using data comprising of examples (labelled data) that have both input and output values (Sammut & Webb, 2017).

Semi-supervised learning. It uses both labelled and unlabelled data to perform a supervised or an unsupervised learning task. This learning is also known as self-training, co-training, learning from the labelled and unlabelled data, or transductive learning (Altınel & Ganiz, 2016; Altınel, Ganiz, & Diri, 2017; Blum & Mitchell, 1998; Muslea, 2010; Rossi, de Andrade Lopes, & Rezende, 2017; Zhu, 2010). Web page classification is widely discussed in the text classification field as a conventional example of this learning method. It uses a co-training schema based on two views, such as text content of a web page including the anchor text of any web page linking to this web page (Blum & Mitchell, 1998).

Ensemble learning. It refers to procedures employed to train multiple classifiers by combining their outputs by considering them as a "committee" of decision makers. Various approaches accomplish this learning concept, for example, bagging, boosting, AdaBoost, stacked generalisation, mixtures of experts, and voting based methods (Brown, 2010; Catal & Nangir, 2017; Elghazel et al., 2016; Lochter et al., 2016; Sammut & Webb, 2010c; Shi, Ma, Xi, Duan, & Zhao, 2011; Ting & Witten, 1997; 1999; Wolpert, 1992; Zelaia, Alegria, Arregi, & Sierra, 2011).

Active learning. This term is used to refer to a learning problem or system where a training algorithm has some role in determining the data it will use for training. In this case, the algorithm is allowed to query a data provider to label an additional subset of training instances (Cohn, 2010; Hu, Mac Namee, & Delany, 2016).

Transfer learning. It refers to the ability of a learning mechanism to improve the performance for a current task after having learned a different but related concept or skill from a previous task. This type of learning is also known as inductive transfer or transfer of knowledge across domains (Do & Ng, 2006; Sun, Feng, & Saenko, 2016; Tan, Zhong, Xiang, & Yang, 2014; Vilalta, Giraud-Carrier, Brazdil, & Soares, 2010; Weiss, Khoshgoftaar, & Wang, 2016).

Multi-view learning. It is also known as data fusion or data integration from multiple feature sets, multiple feature spaces, or diversified feature spaces that may have different distributions of features. This is a case where the data views are conditionally independent sets of the features, and each view has a specific statistical property (Sun, 2013; Woźniak, Graña, & Corchado, 2014; Xu, Tao, & Xu, 2013; Zhao, Xie, Xu, & Sun, 2017). Such learning aims to learn a single function by modelling a particular view, and then to jointly optimise all the functions to improve the generalisation performance (Zhao et al., 2017). We can enumerate articles, for instance Pinheiro et al. (2017), Bhushan and Danti (2017), Lim, Lee, and Kim (2005), Dasigi, Mann, and Protopopescu (2001), related

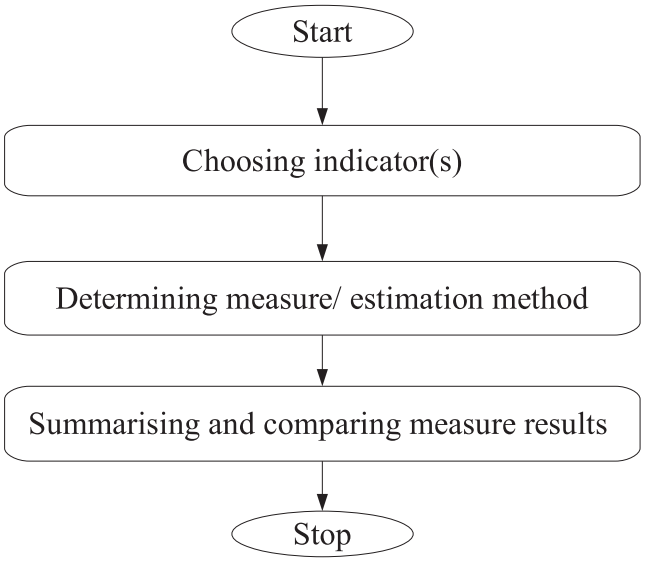

Fig. 5. Basic steps of model evaluation.

to the topic of multi-view text classification. Furthermore, the most similar works in this field focus on semi-supervised learning. This learning utilises the co-training method mentioned earlier to resolve a given multi-view classification task (Gu, Zhu, & Zhang, 2009; Hajmohammadi, Ibrahim, & Selamat, 2014; Matsubara, Monard, & Batista, 2005).

4.2.6. Solution evaluation

Once a classification model is formed, we can choose and establish indicators to measure its performance. Fig. 5 shows the three phases of the model evaluation.

First, we measure a single indicator or group of indicators. There are several state-of-the-art statistical indicators, including precision, recall, accuracy, *F*-score, specificity, area under the curve (AUC), and error rate, etc. (Sokolova & Lapalme, 2009; Vanderlooy & Hüllermeier, 2008). They are described at a macro or micro level (Sokolova & Lapalme, 2009). Their computation procedure is related to the problem of classification tasks, i.e. binary, multi-class, and multi-labelled (Manning et al., 2008; Sokolova & Lapalme, 2009). In addition, we may select more performance-oriented indicators, such as CPU time training, CPU time testing, and memory allocated to the classification model (Ali, Lee, & Chung, 2017).

Second, we establish a method to measure or estimate the indicator values. Subsequently, we determine the method for dividing the available dataset or datasets. For this purpose, procedures, such as leave-one-out and ten-fold cross-validation may be used (Forman & Scholz, 2010; Sammut & Webb, 2010a; Wong, 2015).

Third, all the measured values are summarised and compared. In this phase, we may utilise several methods, such as (1) different plots to visualise and compare the results, e.g. receiver operating characteristic (ROC) (Flach, 2010; Sammut & Webb, 2010b), (2) statistical tests to summarise and examine used processes of text classification (Demsar, 2006; Santafe, Inza, & Lozano, 2015), and (3) a multi-criteria decision making approach (MCDM) to generate the ranking for used solutions (Ali et al., 2017; Kou, Lu, Peng, & Shi, 2012; Krohling, Lourenzutti, & Campos, 2015).

These basic steps produced a well-known experiment plan to evaluate the performance of different methods from text classification. Furthermore, an experiment plan with selected evaluation techniques is strongly related to a chosen research objective.

4.3. Summary of qualitative analysis

We can enumerate the following main conclusions based on qualitative analysis of each stage of text classification process:

- It can be concluded that the issue of text classification occurs in various fields of human activity. There are numerous works explaining various aspects of text classification, such as its domain, tasks, and solutions. Researchers study and develop solutions to resolve general or specific types of problems related to text classification.
- There are numerous well-described and investigated datasets. They can be used to test different techniques for text classification in various applications.
- Multi-instance labelling is not widely discussed and investigated in the research on text classification. Most of the studies are related to labelling of single examples.
- Both vector and graph representation of textual data have been presented in the literature, but among the two the vector representation seems to be more commonly used and discussed in the literature. The feature creation process is not only the simple pre-processing of a text document including stop-words removal with word stemming or lemmatisation, but also a sophisticated process utilising knowledge from different areas, such as information theory, linear algebra, statistics, and natural language processing (NLP). The most recent feature weighting methods outperform the standard and well-known weighting functions.
- The issue of feature selection is widely and thoroughly discussed in the literature. A significant number of publications indicate that this topic is well explored in the context of text classification. The PCA, SVD, and LDA algorithms of feature projection for text classification are discussed the most frequently. We have not found any compelling work related to the investigation of nonlinear PCA, Sammon, tSNE, or CSPCA techniques for document categorisation.
- Numerous works related to supervised learning, seems to be deeply explored. The semi-supervised approach is closely related to the multi-view learning technique. In some cases, it utilises the idea of multi-views in a classification process, for example, Hajmohammadi et al. (2014) developed a multi-view semi-supervised learning system that resolves a cross-lingual sentiment classification problem. The ensemble learning method is also closely related to the multi-view learning approach. From one perspective, it is a specific case of multi-view learning, where the techniques to create data views are utilised, i.e. partitioning a feature space horizontally or vertically. Consequently, the classifiers committee utilises the data views. However, we did not encounter any work that could describe and utilise the partitioning approach in the context of text classification. Some studies propose untypical solutions, for example, Tan et al. (2014) developed a multi-transfer solution which is based on transfer learning with multiple views and multiple sources.
- Although the evaluation procedure is rather general, its implementation is strictly related to a document classification objective. There are numerous indicators and different methods to measure their values. The most prominent indicators are precision, recall, and *F*-measure. Usually, they are measured using the ten-fold cross-validation technique.

5. Quantitative analysis of studies

In this section, we describe the quantitative analysis conducted of the studies related to text classification.

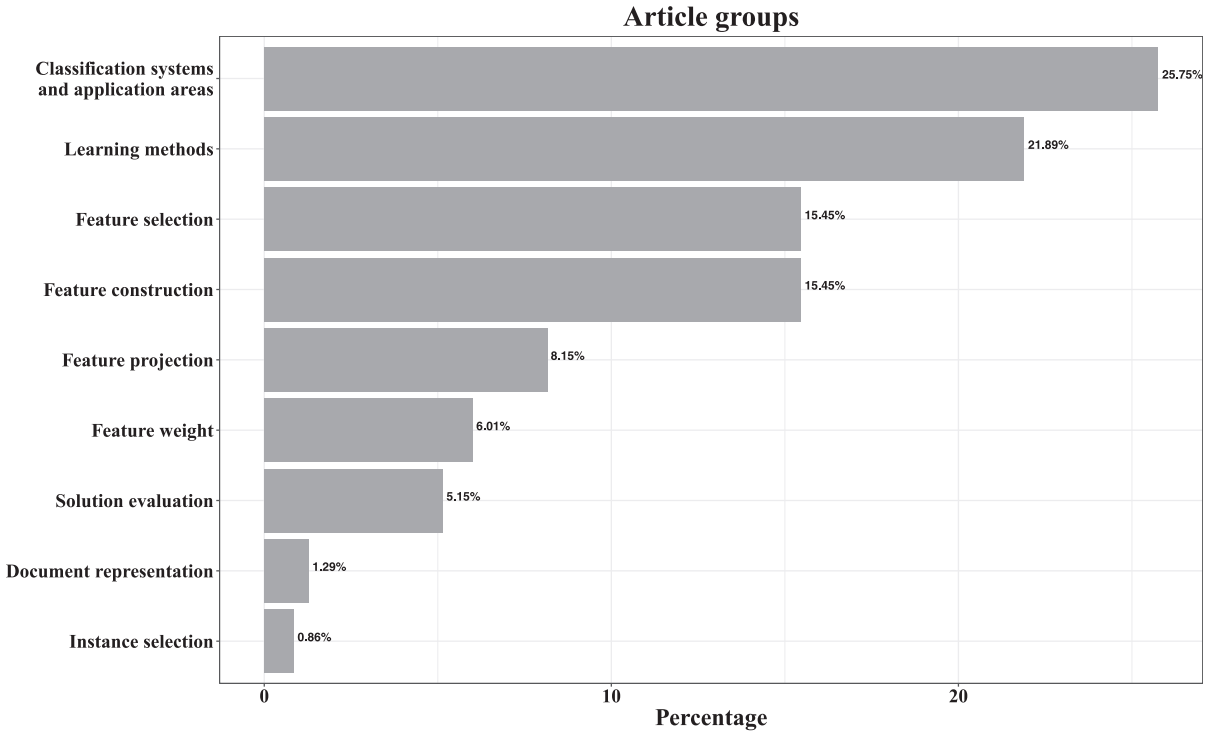

Fig. 6. Distribution of the articles over selected topics, i.e. *classification systems and application areas, document representation, feature construction, feature weight, feature selection, feature projection, instance selection, learning methods*, and *evaluation methods*. The research sample contains 233 articles.

5.1. Research questions

Because journals including conferences are platforms where scientists share and discuss their results, so that the quantitative analysis, is a type of analysis of research forums. Such analysis provides the opportunity to display a quantitative map of the research topics. More specifically, it assists us in obtaining the answers for the following research questions:

1. Which parts of the text classification framework are investigated the most and least?
2. What does a distribution of the articles look like?
3. What does a number of articles in each topic with the change of the years?
4. What does a distribution of the articles in journals resembles?
5. Can we highlight the top journals in text classification?
6. Can we highlight the top countries that publish extensively in text classification?
7. How does look like cooperation between countries?
8. What does a distribution of the number of authors in publications look like?

We attempt to provide answers to the above questions by qualitatively analysing our dataset that contains 242 selected articles, as presented in Section 3.2. For this purpose, we present and discuss the distribution of manuscripts by their research topics, publication time, and journals.

5.2. Research topics

The articles are grouped into nine categories as presented in Fig. 6. It includes a bar chart showing their percentage share in the entire dataset. We analyse 233 articles, the omitted 9 articles are reviews or unclassified materials. Each category in Fig. 6 consists the following number of articles, respectively:

- Classification systems and application areas (60 articles);
- Learning methods (51 articles) – it is the sum of Labelling methods (8 articles) and Learning methods (43 articles), if we take into account the classification in Table 2;
- Feature construction (36 articles) – it is the subtraction of Document representation (3 articles) from Feature construction (39 articles), if we take into account the classification in Table 2;
- Feature selection (36 articles) – it is the subtraction of Instance selection (2 articles) from Feature selection (38 articles), if we take into account the classification in Table 2;
- Feature projection (19 articles);
- Solution evaluation (12 articles).

Based on Fig. 6, it can be easily noticed that the most prominent research topic in text classification is related to features selection, construction, weighting, and projection. Over 2/5ths of the studies deal with these issues, and approximately 1/5th of the papers are dedicated to learning methods, which shows the distinctive focus of the researchers on this problem. The same can be stated for various approaches for the construction of classification systems including their application areas because 1/4th of the works deal with this topic. The remaining issues, namely, evaluation methods, document representation, and instance selection are found in 1/5th of the studies. Among this group, only evaluation methods have a considerable contribution .

In Fig. 7 we present the distribution of articles in classification research topics over publication years. Some trends can be distinguished if we scrutinise the number of articles in classification research topics over publication years. In the recent years, one may notice the predominant interest in feature construction and selection; learning methods; and classification systems and application areas. Among them, two topics, i.e. (i) learning methods and (ii) classification systems and application areas seem to be worth exploring for scientists despite the course of years. Conversely, feature projection gains less attention currently than before. The remaining categories do not show observable patterns.

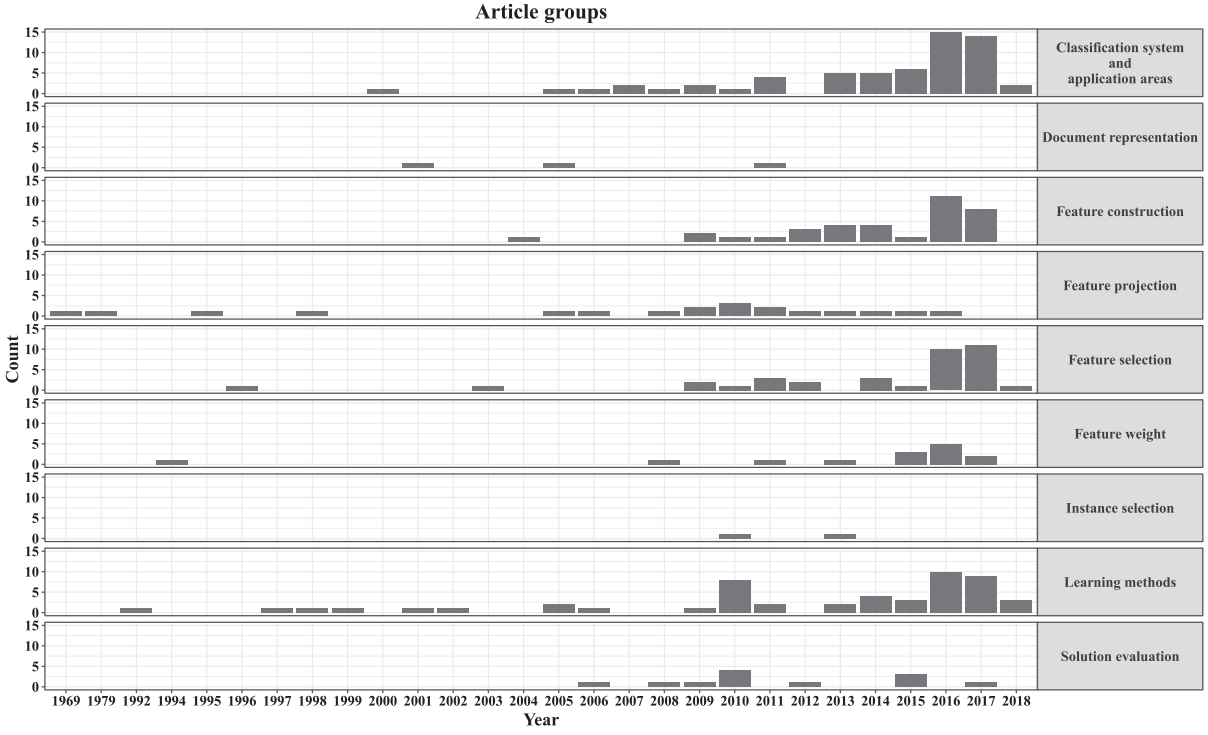

Fig. 7. Distribution of articles in classification research topics over publication years. The research sample contains 233 articles.

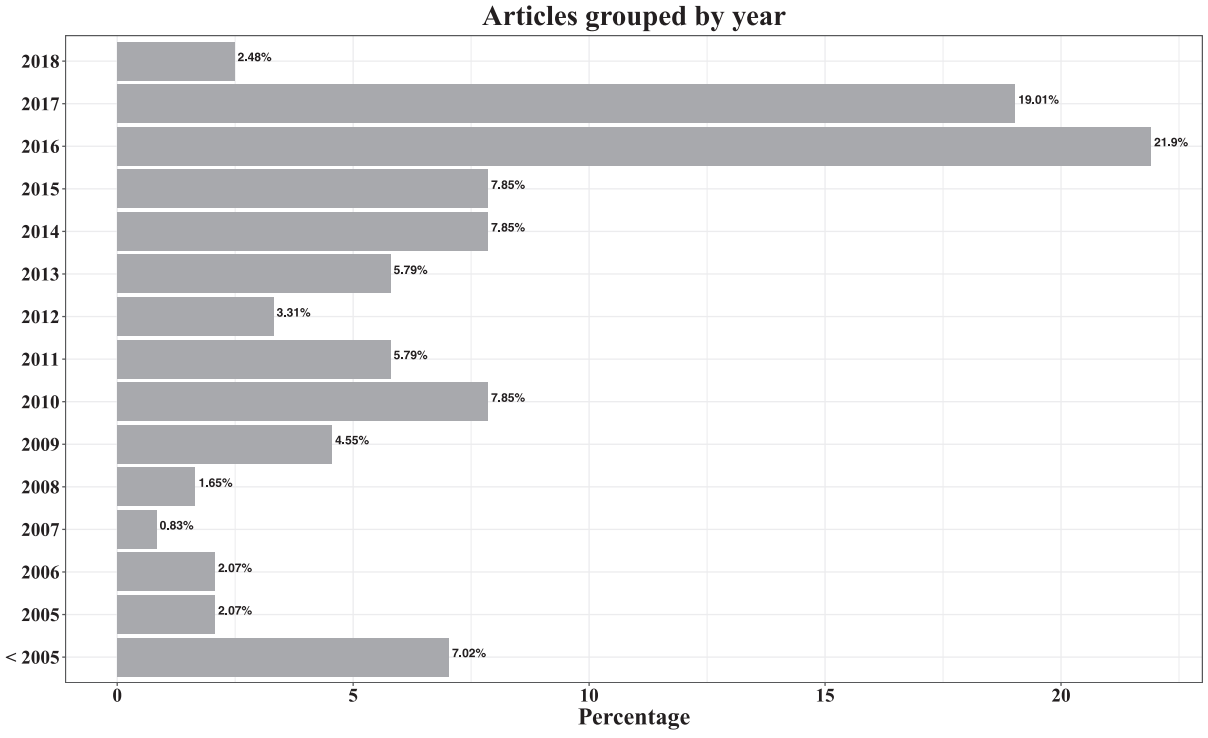

Fig. 8. Distribution of articles based on publication years. There are fourteen categories; thirteen among them represent years from 2005 to 2018. Moreover, one category contains the works published before 2005. The research sample contains 242 articles.

5.3. Publication time

All the 242 articles from our dataset are grouped into fifteen categories. Each category represents a publication year. The papers published earlier than 2005 are grouped into a category called ≤ 2005. Fig. 8 presents a bar chart of the distribution of articles by publication time.

In Fig. 8, 2/5th of the articles are published in 2016 and 2017. A continuous decrease in the number of published works can be observed from 2015 onward. However, there are some irregularities, such as a more significant decrease in 2012 or sudden increase in 2010 and 2006. It is apparent that in recent years, researchers pay more attention to text classification. One-fifth of the manuscripts are from 2017. Simultaneously, 2016 is the year the

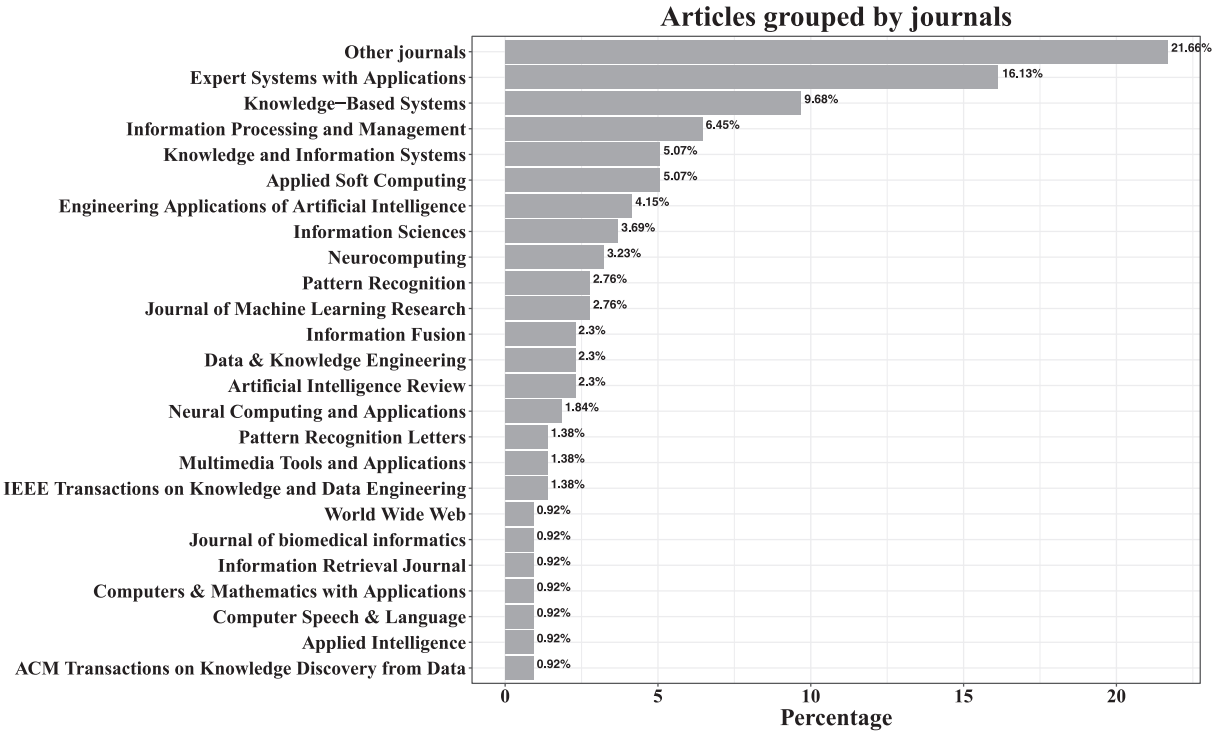

Fig. 9. Distribution of articles by journals including conference proceedings. The research sample contains 217 articles.

most prolific for the publications. Moreover, it is highly distinctive that the number of works doubles between 2016 and 2017. One-fifth of the papers appear before 2005. Since 2005 we observe almost a linear increase in the number of works each year until its profound increase in 2016. Accordingly, we can state that text classification is a well-known issue that has a long tradition of research.

5.4. Journals

In the next part of our quantitative assessment, we systematise the articles into twenty five categories. Each category represents the different journals in which the articles are published. The journals or conference series containing only one article are grouped into one category named *Other journals*. The *Other journals* category includes journals such as *Pattern Analysis and Applications, Computational Statistics and Data Analysis, Journal of Applied Logic, Statistical Analysis and Data Mining, Journal of Systems and Software, Machine Learning, ACM Computing Surveys, Transactions on Information Systems*. The *Other journals* category also includes conference proceedings, such as the *International Joint Conference on Artificial Intelligence, International Conference on Advanced Data Mining and Applications, ACM International Conference on Information and Knowledge Management, International Conference on Computational Learning Theory*, and *AAAI Conference on Artificial Intelligence*. The research sample contains 217 articles because the 25 articles published in book chapters were excluded from our dataset. Fig. 9 presents a bar chart of the article distribution by journals including conference proceedings.

Fig. 9shows that there are three most distinctive journals in text classification, i.e. *Expert System with Application, Knowledge Based-System*, and *Information Processing and Management*. They cover approximately 1/3rd of the publications from our dataset. In contrast, 1/3rd of the works originate from journals or conference proceedings with only one publication in the dataset. The distribution shows a significant dispersion of the research discussion forums in the text classification. In between these two extreme observations, there are several journals including conference proceedings, each containing between 1% and 4% of the works. Together they cover over 1/3rd of the studies.

5.5. Distribution of publications by countries and a cooperation network

We were able to identify country names in 218 publications from our database. Only 24 papers without a country name we had to exclude from analysis. We have to note that a country name means an affiliation country of an author. An extraction process produced 690 different affiliations with many repetitions of a country name. As a result, we identified 42 unique country names. Figs. 10 and 11 show distribution of publications by countries and a cooperation network between them, respectively.

The percentage distribution of authors by their affiliation countries is shown in Fig. 10. This statistic also reflects the number of papers in text classification by countries. However, it does it only to some extent because these works may have many co-authors affiliated to various countries. Based on Fig. 10, we can conclude that nowadays the superior position in text classification takes China, as 1/4th of authors comes for this country. The second country is The United States having 1/8 of authors in this research area. The third position goes to Brazil. We must underline that these numbers say nothing about the quality of studies. They cannot show in which country appeared the breakthroughs in text classification.

An interesting issue is a cooperation across countries in research on text classification. Papers usually have more than one author, and they may originate from different countries. Thus, we can easily produce a diagram showing a cooperation network between countries by analysing affiliation countries of authors (Fig. 11). By scrutinising this network diagram, we can distinguish three types authors. The first group forms researchers who work only with co-workers for their country. In Fig. 11, these countries lay in the outer semi-circle. The second group comprises scientists

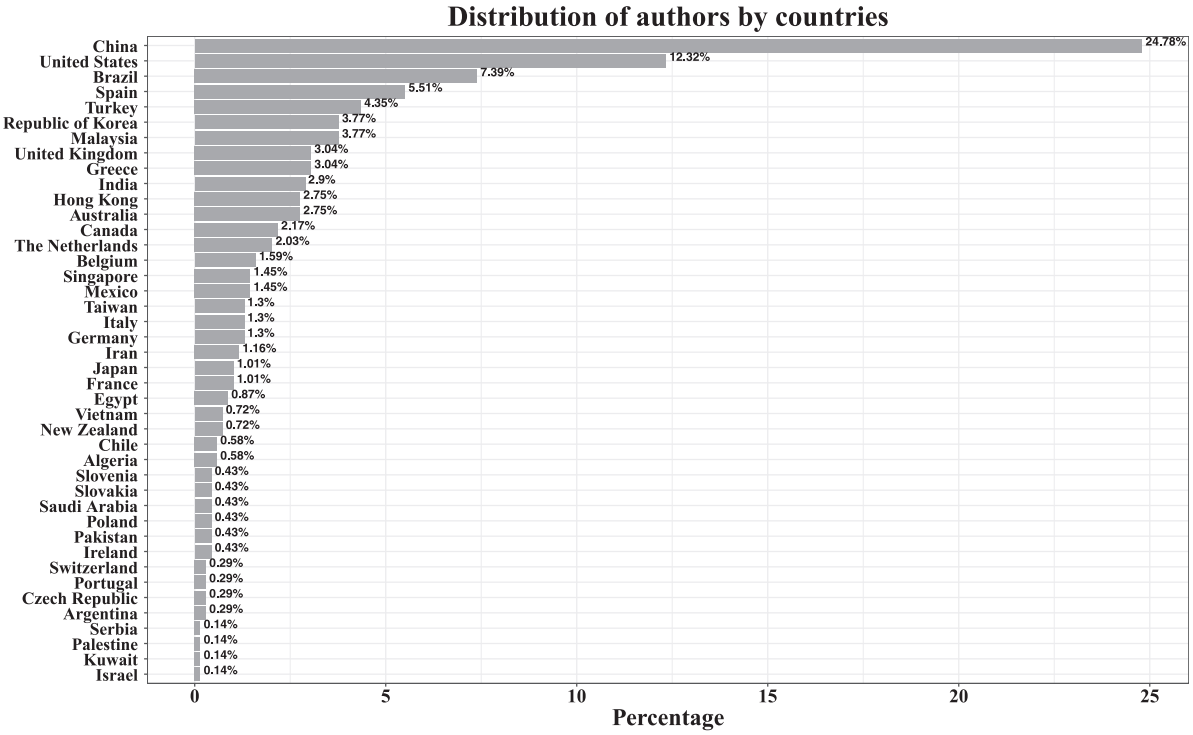

Fig. 10. Distribution of authors by their affiliation countries. The research sample contains 218 articles, which produced 690 different affiliations to 42 unique country names.

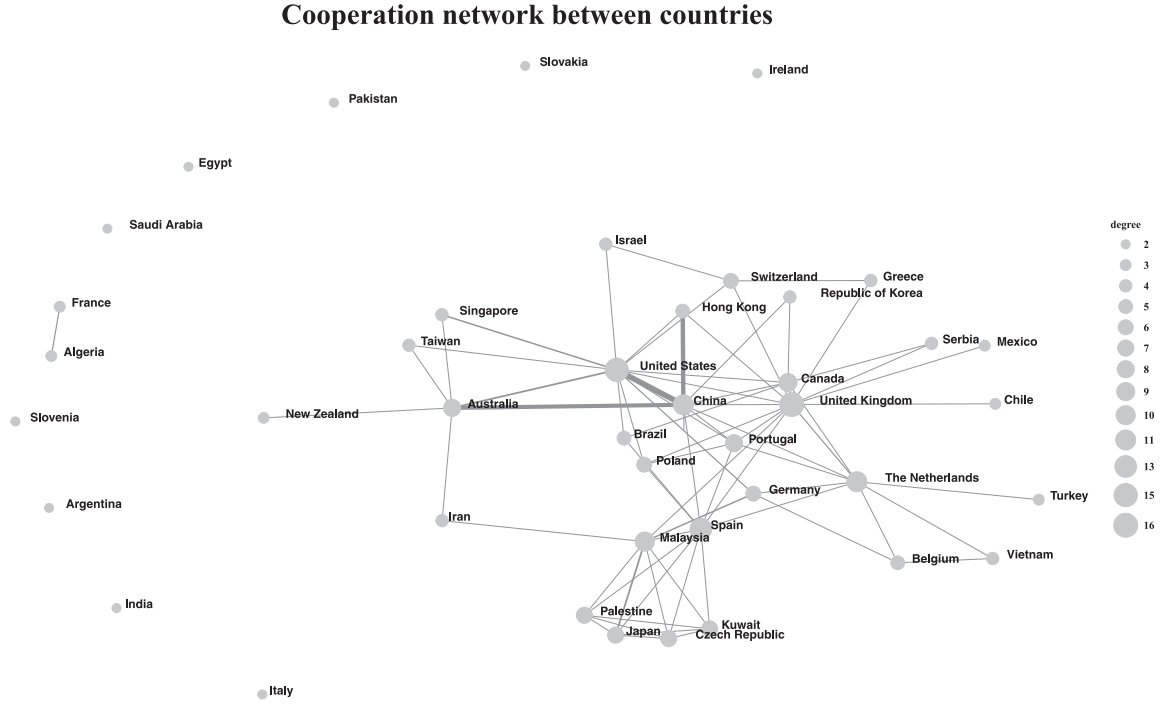

Fig. 11. A cooperation network between countries. The research sample contains 218 articles, which produced 690 different affiliations to 42 unique country names.

who publish with at least one co-author from another country. These countries are mainly located in the outer sphere of the most prominent cooperation cluster in Fig. 11. The third group are scientists who worked on papers with co-authors originating from more than two countries. These countries are mainly laid in the centre of the cluster in Fig. 11. Unsurprisingly, they correspond to the countries with the highest number of publications as shown in Fig. 10. The thicker a line connecting countries is, the more co-work has been noticed. The most distinguishing are links between China (including Hong Kong) and The United States; and China and Australia. However, it impossible to find any pattern linking scientists based on their geographical location or known similarities between countries, or historical issues. It seems that researchers ignore political and historical conditions and focus only scientific issues.

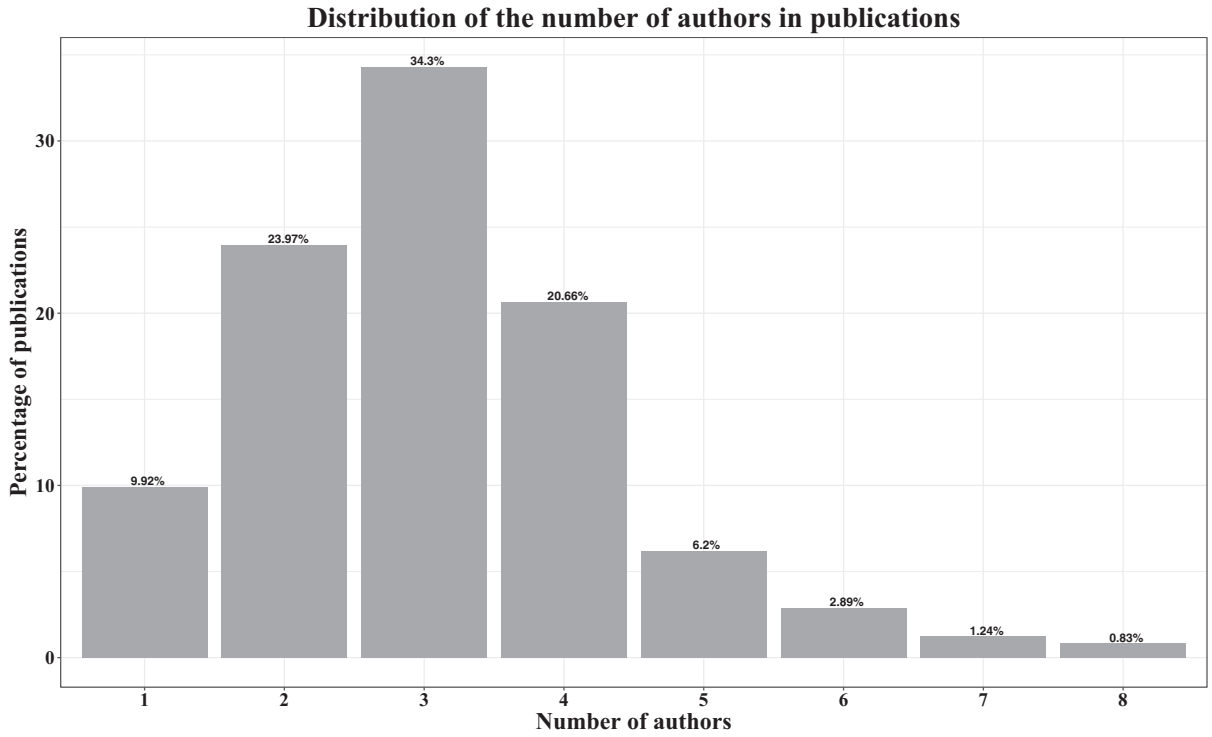

Fig. 12. Distribution of the number of authors in publications. The research sample contains 242 articles, which produced 743 different names of authors with a repetitions.

5.6. Distribution of the number of authors in publications

The 242 articles from our dataset are analysed in term of author numbers. We have 743 different names of authors with repetitions. Fig. 12 presents a bar chart of the distribution of the number of authors in publications.

Fig. 12 shows how many authors usually have papers on text classification. About 1/3rd of studies are prepared in the team of three people. Whereas, 1/4th and 1/5th of works contain two and four authors respectively. As we can notice, the majority (3/4th) of studies are carried out in teams composed of 2–4 researchers. Only about 1/10th of papers have a sole author and 1/15th are prepared in teams of four or more people.

5.7. Summary of quantitative analysis

At the beginning of this section, we posed several research question. Herein, we answer these question by summarising the quantitative analysis as follows:

- Which parts of the text classification framework are investigated the most and least?
  The most text investigated parts of the classification framework are the issues regarding feature selection, construction, weighting, and projection, as well as training or construction of models. The least mentioned topic is the evaluation method. However, it is difficult to establish whether this implies that this topic should be explored more, or whether it is perceived as unimportant, or conclude as if nothing can be done.
- What does a distribution of the articles look like?
  Text classification problem has been investigated for a long time. The interest of researchers in it continually increases over the years. Currently, this is a prominent topic.
- What does a number of articles in each topic with the change of the years? What does a distribution of the articles in journals resembles? Can we highlight the top journals in text classification?
  We observe almost a linear increase in the number of works over the course of years, including the sudden increase in the recent years. The quantitative analysis shows that the journals including conference proceedings in text classification can be categorized into three groups, namely, (1) most popular journals, e.g. *Expert System with Application, Knowledge Based-System*, and *Information Processing and Management*, (2) journals including conference proceedings with moderate and uniform share, and (3) remaining periodicals, where text classification studies appear only occasionally. Each group includes about 1/3rd of the papers from our dataset.
- Can we highlight the top countries that publish extensively in text classification? How does look like cooperation between countries? What does a distribution of the number of authors in publications look like?
  We observed three groups of authors, namely: (i) researchers publishing only with colleagues form their country; (ii) scientists cooperating with other countries; and (iii) teams extensively co-working over the world. The last group has the highest number of publications. Most publications have between 2 and 4 of authors.

Additional quality remarks. Despite the quantitative analysis, we express our subjective view on the quality of these sources. The option is based on our overall impression acquired during the literature review. In our view, the crucial and relevant sources are the *Advances in Data Analysis and Classification, Journal of Classification, Annals of Mathematics and Artificial Intelligence*, Springer computer science series (for example, *Lecture Notes in Computer Science, Communications in Computer and Information Science,Theory and Applications of Natural Language Processing, Studies in Computational Intelligence*), and different journals published by ACM or IEEE. Furthermore, we can enumerate some top conferences excluded from our bibliography but which have a significant effect on an investigated area. The main conferences are the following: *European Conference on Machine Learning and Principles and Practice of*

Knowledge Discovery in Databases, ACM series of international conferences called *Web Search and Data Mining* or *SIGKDD: The Community for Data Mining, Data Science and Analytics*. In addition, the *Very Large Data Base* conference series may be useful to identify new concepts.

6. Conclusions

In this study, we reviewed the works dealing with text classification. Accordingly, we wanted to achieve three objectives, namely, (1) to extract the most crucial phases of the text classification process, (2) to qualitatively analyse each phase to identify the common techniques, and (3) to quantitatively analyse studies to observe some trends.

By realising the first objective, we identified the most distinctive phases in the process of text classification, which we called the text classification framework. The framework included several elements, namely, (1) data acquisition, (2) data analysis and labelling, (3) feature construction and weighting, (4) feature selection and projection, (5) training of a classification model, and (6) solution evaluation. In addition, we constructed a vocabulary allowing us to query search engines for papers corresponding to various sub-fields of text classification. The framework and vocabulary together constitute a map of this research topic. The map should assist researchers for promptly addressing their questions and realise the complexity of the topic.

On accomplishing the second objective, we examined the literature related to each phase of text classification. The studies were based on applying algorithms in particular domains. Alternatively, they deal with the methods development in a more general way. Researchers published several well-described and highly tested datasets available to everyone for testing new approaches. Most of them were labelled for supervised learning. The studies usually utilised simple data examples. Multi-instance labelling was rather uncommon. Textual data could be represented in a vector or graph form, with the first form being the most prominent. We have to highlight that document representation is the vital issue for classification quality. There are some recent methods for feature construction, e.g. Word2vec, GloVe, and their modifications that outperform the classification results in some cases (Araque, Corcuera-Platas, Sánchez-Rada, & Iglesias, 2017; Kamkarhaghighi & Makrehchi, 2017; Li, Li, Fu, Masud, & Huang, 2016). Some authors advocated that it is advisable to reduce the feature dimensionality. For this purpose, they applied the PCA, SVD, and LDA algorithms the most frequently. Various algorithms can be used for classifiers training. They can be grouped into several types, including supervised, semi-supervised, ensemble, active, transfer and multi-view learning. The most frequently explored approach was the supervised learning. Trained models were usually assessed by using a ten-fold cross-validation procedure with quality indicators such as precision, recall and *F*-score. However, some studies used more sophisticated measures.

The third objective involved the quantitative analysis of the studies based on the research topics, publication date, and publication location. We found that in recent years there has been an observable increase in the number of papers on text classification. In the group of journals, including *Expert System with Application, Knowledge Based-System*, and *Information Processing and Management*, the discussion regarding this topic was notably lively; whereas, others journals were not particularly focused on it. It was noted that researchers primarily paid significant attention to the issues of features selection, construction, weighting, and projection, as well as to the training or construction of models.

It can be concluded that text classification is a well-developed research topic. Simultaneously, it is also a prominent subject in which new approaches can be discovered and the findings can be utilised in various domains. Several issues have not been thoroughly addressed yet. We did not find works explicitly related the problems of over-fitting of text classification models, or transfer, multi-view learning, and dynamic selection classifier, which is the most promising approach for multiple-classifier systems (Cruz, Sabourin, & Cavalcanti, 2018). Moreover, the concept drift which is strong related to data stream analysis (Krawczyk, Minku, Gama, Stefanowski, & Woźniak, 2017) requires more research attention.

In our opinion, the current hottest topics are multi-lingual or cross-lingual text classification (García, Rodríguez, & Rifón, 2017; Wei, Lin, & Yang, 2011; Zhou, Zhu, He, & Hu, 2016), text stream analysis (Katakis, Tsoumakas, & Vlahavas, 2010; Nanculef, Flaounas, & Cristianini, 2014; Yang, Zhang, & Li, 2011; Zhang, Chu, Li, Hu, & Wu, 2017), opinions or sentiments analysis (Araque et al., 2017; García-Pablos, Cuadros, & Rigau, 2018; Kang, Ahn, & Lee, 2018; Sun, Luo, & Chen, 2017; Tripathy, Anand, & Rath, 2017), and ensemble learning method (Cruz et al., 2018). Furthermore, many works are related to embedding features to create better semantics vocabularies for classifiers learning (Araque et al., 2017; Kamkarhaghighi & Makrehchi, 2017; Li et al., 2016; Sun et al., 2017).

Because our study has some limitations, there is scope for further research. First, we omitted a description of the machine learning algorithms because this topic has been widely discussed in other literature reviews. However, a new systematic comparison of the algorithms may be relevant to some readers. Second, it would be worth exploring additional publication sources, including ACM and IEEE. We omitted these databases owing to our licence limitation. Also, we excluded arXiv since it mostly includes pre-prints which are only moderated but not peer-reviewed. Third, currently, text streaming is an emerging topic worth analysing. More information on it can be found in Nguyen, Woon, and Ng (2015), Nanculef et al. (2014). In addition, classification techniques other than machine learning has been considered in the following reviews. For example, some researchers, constructed classification models based on fuzzy logic which were more interpretable than those based on machine learning (Sharef & Martin, 2015). Finally, Sun et al. (2017), Aggarwal and Zhai (2012) enumerated several tools for text classification, namely, OpenNLP and Standford solutions. More recently, tools based on Spark, Python, or R-project have also been developed. It appears to be a good idea to present all such tools in a systematic manner.

While being aware of the advantages and disadvantages of our study, we believe that it will assist other scientists and professionals to conduct and publish their studies in the area of text classification.

Acknowledgements

The authors would like to thank Jakub Kierzkowski for his consultation for the construction of the mathematical framework for text classification. Because of the generated notation, we better understood the entire system and clarified our approach for text classification. Moreover, we would like to thank the editor as well as the reviewers for the time and effort that they have dedicated to our work.

References

Aas, K., & Eikvil, L. (1999). Text categorisation: A survey. *Technical report*. Norwegian Computing Center.

Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems, 26*(3), 12.

Adeva, J. G., Atxa, J. P., Carrillo, M. U., & Zengotitabengoa, E. A. (2014). Automatic text classification to support systematic reviews in medicine. *Expert Systems with Applications, 41*(4, Part 1), 1498–1508.

Agathangelou, P., Katakis, I., Koutoulakis, I., Kokkoras, F., & Gunopulos, D. (2018). Learning patterns for discovering domain-oriented opinion words. *Knowledge and Information Systems, 55*(1), 45–77.

Aggarwal, C. C. (2015). *Data mining: The textbook*. Springer.

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In C. C. Aggarwal, & C. Zhai (Eds.), *Mining Text Data* (pp. 163–222). Springer 10.1007/978-1-4614-3223-4_6.

Agnihotri, D., Verma, K., & Tripathi, P. (2017). Variable global feature selection scheme for automatic classification of text documents. *Expert Systems with Applications, 81*, 268–281.

Al-Salemi, B., Noah, S. A. M., & Ab Aziz, M. J. (2016). RFBoost: An improved multi-label boosting algorithm and its application to text categorisation. *Knowledge-Based Systems, 103*, 104–117.

Alalga, A., Benabdeslem, K., & Taleb, N. (2016). Soft-constrained Laplacian score for semi-supervised multi-label feature selection. *Knowledge and Information Systems, 47*(1), 75–98. doi:10.1007/s10115-015-0841-8.

Ali, R., Lee, S., & Chung, T. C. (2017). Accurate multi-criteria decision making methodology for recommending machine learning algorithm. *Expert Systems with Applications, 71*, 257–278. doi:10.1016/j.eswa.2016.11.034.

Almeida, T. A., Silva, T. P., Santos, I., & Hidalgo, J. M. G. (2016). Text normalization and semantic indexing to enhance instant messaging and sms spam filtering. *Knowledge-Based Systems, 108*, 25–32.

Alpaydın, E., Cheplygina, V., Loog, M., & Tax, D. M. (2015). Single-vs. multiple-instance classification. *Pattern Recognition, 48*(9), 2831–2838.

Altınçay, H., & Erenel, Z. (2014). Ternary encoding based feature extraction for binary text classification. *Applied Intelligence, 41*(1), 310–326.

Altınel, B., & Ganiz, M. C. (2016). A new hybrid semi-supervised algorithm for text classification with class-based semantics. *Knowledge-Based Systems, 108*, 50–64.

Altınel, B., Ganiz, M. C., & Diri, B. (2015). A corpus-based semantic kernel for text classification by using meaning values of terms. *Engineering Applications of Artificial Intelligence, 43*, 54–66.

Altınel, B., Ganiz, M. C., & Diri, B. (2017). Instance labeling in semi-supervised learning with meaning values of words. *Engineering Applications of Artificial Intelligence, 62*, 152–163.

Araque, O., Corcuera-Platas, I., Sánchez-Rada, J. F., & Iglesias, C. A. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications, 77*, 236–246.

Armanfard, N., Reilly, J. P., & Komeili, M. (2016). Local feature selection for data classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(6), 1217–1227. doi:10.1109/TPAMI.2015.2478471.

Badawi, D., & Altinçay, H. (2014). A novel framework for termset selection and weighting in binary text classification. *Engineering Applications of Artificial Intelligence, 35*, 38–53. doi:10.1016/j.engappai.2014.06.012.

Bandhakavi, A., Wiratunga, N., Padmanabhan, D., & Massie, S. (2017). Lexicon based feature extraction for emotion text classification. *Pattern Recognition Letters, 93*, 133–142. doi:10.1016/j.patrec.2016.12.009.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the fifty-second annual meeting of the association for computational linguistics: 1* (pp. 238–247). (Long papers)

Basto-Fernandes, V., Yevseyeva, I., Méndez, J. R., Zhao, J., Fdez-Riverola, F., & Emmerich, M. T. (2016). A spam filtering multi-objective optimization study covering parsimony maximization and three-way classification. *Applied Soft Computing, 48*, 111–123.

Bellotti, T., Nouretdinov, I., Yang, M., & Gammerman, A. (2014). Feature selection. In V. Balasubramanian, S.-S. Ho, & V. Vovk (Eds.), *Conformal prediction for reliable machine learning* (pp. 115–130)). Elsevier.

Bhushan, S. N. B., & Danti, A. (2017). Classification of text documents based on score level fusion approach. *Pattern Recognition Letters, 94*, 118–126. doi:10.1016/j.patrec.2017.05.003.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on computational learning theory* (pp. 92–100). ACM.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the biennial GSCL conference on from form to meaning: Processing texts automatically, 2009* (pp. 31–40). Tübingen. (vol. Normalized

Brown, G. (2010). *Encyclopedia of machine learning* (pp. 312–320). Boston, MA: Springer US.

Bui, D. D. A., Del Fiol, G., & Jonnalagadda, S. (2016). PDF text classification to leverage information extraction from publication reports. *Journal of Biomedical Informatics, 61*, 141–148.

Burges, C. J. C. (2010). Dimension reduction: A guided tour. *Foundations and Trends in Machine Learning, 2*(4), 275–365. doi:10.1561/2200000002.

Cagliero, L., & Garza, P. (2013). Improving classification models with taxonomy information. *Data & Knowledge Engineering, 86*, 85–101. doi:10.1016/j.datak.2013.01.005.

Castro, D., Souza, E., Vitório, D., Santos, D., & Oliveira, A. L. I. (2017). Smoothed n-gram based models for tweet language identification: A case study of the brazilian and european portuguese national varieties. *Applied Soft Computing, 61*, 1160–1172. doi:10.1016/j.asoc.2017.05.065.

Catal, C., & Nangir, M. (2017). A sentiment classification model based on multiple classifiers. *Applied Soft Computing, 50*, 135–141.

Chang, M., & Poon, C. K. (2009). Using phrases as features in email classification. *Journal of Systems and Software, 82*(6), 1036–1045. doi:10.1016/j.jss.2009.01.013.

Chang, X., Nie, F., Yang, Y., Zhang, C., & Huang, H. (2016). Convex sparse PCA for unsupervised feature learning. *ACM Transactions on Knowledge Discovery from Data, 11*(1), 3.

Charalampakis, B., Spathis, D., Kouslis, E., & Kermanidis, K. (2016). A comparison between semi-supervised and supervised text mining techniques on detecting irony in greek political tweets. *Engineering Applications of Artificial Intelligence, 51*, 50–57.

Chaturvedi, I., Ong, Y., Tsang, I. W., Welsch, R. E., & Cambria, E. (2016). Learning word dependencies in text by means of a deep recurrent belief network. *Knowledge-Based Systems, 108*, 144–154. doi:10.1016/j.knosys.2016.07.019.

Chen, K., Zhang, Z., Long, J., & Zhang, H. (2016). Turning from TF-IDF to TF-IGM for term weighting in text classification. *Expert Systems with Applications, 66*, 245–260. doi:10.1016/j.eswa.2016.09.009.

Chen, L., Guo, G., & Wang, K. (2011). Class-dependent projection based method for text categorization. *Pattern Recognition Letters, 32*(10), 1493–1501. doi:10.1016/j.patrec.2011.01.018.

Choi, D., Ko, B., Kim, H., & Kim, P. (2014). Text analysis for detecting terrorism-related articles on the web. *Journal of Network and Computer Applications, 38*, 16–21. doi:10.1016/j.jnca.2013.05.007.

Cohn, D. (2010). Active learning. In C. Sammut, & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 10–14). Boston, MA: Springer US. doi:10.1007/978-0-387-30164-8_6.

Cruz, R. M., Sabourin, R., & Cavalcanti, G. D. (2018). Dynamic classifier selection: Recent advances and perspectives. *Information Fusion, 41*, 195–216.

Cunningham, J. P., & Ghahramani, Z. (2015). Linear dimensionality reduction: Survey, insights, and generalizations.. *Journal of Machine Learning Research, 16*(1), 2859–2900.

Cuzzola, J., Jovanović, J., Bagheri, E., & Gašević, D. (2015). Automated classification and localization of daily deal content from the web. *Applied Soft Computing, 31*, 241–256.

Dasigi, V., Mann, R. C., & Protopopescu, V. A. (2001). Information fusion for text classification - An experimental comparison. *Pattern Recognition, 34*(12), 2413–2425. doi:10.1016/S0031-3203(00)00171-0.

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, 7*, 1–30.

Diab, D. M., & El Hindi, K. M. (2017). Using differential evolution for fine tuning Naïve Bayesian classifiers and its application for text classification. *Applied Soft Computing, 54*, 183–199.

Do, C. B., & Ng, A. Y. (2006). Transfer learning for text classification. In *Proceedings of the 2006 advances in neural information processing systems* (pp. 299–306).

Du, Y., Liu, J., Ke, W., & Gong, X. (2018). Hierarchy construction and text classification based on the relaxation strategy and least information model. *Expert Systems with Applications, 100*, 157–164.

Duin, R. P., Loog, M., Pękalska, E., & Tax, D. M. (2010). Feature-based dissimilarity space classification. In *Proceedings of the 2010 recognizing patterns in signals, speech, images and videos* (pp. 46–55). Springer.

Elghazel, H., Aussem, A., Gharroudi, O., & Saadaoui, W. (2016). Ensemble multi-label text categorization based on rotation forest and latent semantic indexing. *Expert Systems with Applications, 57*, 1–11. doi:10.1016/j.eswa.2016.03.041.

Enríquez, F., Troyano, J. A., & López-Solaz, T. (2016). An approach to the use of word embeddings in an opinion classification task. *Expert Systems with Applications, 66*, 1–6. doi:10.1016/j.eswa.2016.09.005.

Escalante, H. J., García-Limón, M. A., Morales-Reyes, A., Graff, M., Montes-y-Gómez, M., Morales, E. F., et al. (2015). Term-weighting learning via genetic programming for text classification. *Knowledge-Based Systems, 83*, 176–189. doi:10.1016/j.knosys.2015.03.025.

Fattah, M. A. (2015). New term weighting schemes with combination of multiple classifiers for sentiment analysis. *Neurocomputing, 167*, 434–442. doi:10.1016/j.neucom.2015.04.051.

Figueiredo, F., da Rocha, L. C., Couto, T., Salles, T., Gonçalves, M. A., & Meira Jr. , W. (2011). Word co-occurrence features for text classification. *Information Sciences, 36*(5), 843–858. doi:10.1016/j.is.2011.02.002.

Flach, P. A. (2010). ROC analysis. In C. Sammut, & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 869–875)). Boston, MA: Springer US. doi:10.1007/978-0-387-30164-8_733.

Forman, G., & Scholz, M. (2010). Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter, 12*(1), 49–57. doi:10.1145/1882471.1882479.

de Fortuny, E. J., Smedt, T. D., Martens, D., & Daelemans, W. (2014). Evaluating and understanding text-based stock price prediction models. *Information Processing and Management, 50*(2), 426–441. doi:10.1016/j.ipm.2013.12.002.

Foulds, J., & Frank, E. (2010). A review of multi-instance learning assumptions. *The Knowledge Engineering Review, 25*(01), 1–25.

García, M. A. M., Rodríguez, R. P., & Rifón, L. A. (2017). Wikipedia-based cross-language text classification. *Information Sciences, 406*, 12–28.

García-Pablos, A., Cuadros, M., & Rigau, G. (2018). W2VLDA: Almost unsupervised system for aspect based sentiment analysis. *Expert Systems with Applications, 91*, 127–137.

Ghareb, A. S., Bakar, A. A., & Hamdan, A. R. (2016). Hybrid feature selection based on enhanced genetic algorithm for text categorization. *Expert Systems with Applications, 49*, 31–47.

Giachanou, A., Salampasis, M., & Paltoglou, G. (2015). Multilayer source selection as a tool for supporting patent search and classification. *Information Retrieval Journal, 18*(6), 559–585. doi:10.1007/s10791-015-9270-2.

Giatsoglou, M., Vozalis, M. G., Diamantaras, K., Vakali, A., Sarigiannidis, G., & Chatzisavvas, K. C. (2017). Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications, 69*, 214–224.

Gollapalli, S. D., Caragea, C., Mitra, P., & Giles, C. L. (2013). Researcher homepage classification using unlabeled data. In *Proceedings of the twenty-second interna-*

tional conference on world wide web, WWW '13 (pp. 471–482). New York, NY, USA: ACM. doi:10.1145/2488388.2488430.
Gollapalli, S. D., Caragea, C., Mitra, P., & Giles, C. L. (2015). Improving researcher homepage classification with unlabeled data. *ACM Transactions on the Web, 9*(4), 17:1–17:32. doi:10.1145/2767135.
Gollapalli, S. D., Giles, C. L., Mitra, P., & Caragea, C. (2011). On identifying academic homepages for digital libraries. In *Proceedings of the eleventh annual international ACM/IEEE joint conference on digital libraries, JCDL '11* (pp. 123–132). New York, NY, USA: ACM. doi:10.1145/1998076.1998099.
Gomez, J. C., & Moens, M.-F. (2012). PCA document reconstruction for email classification. *Computational Statistics & Data Analysis, 56*(3), 741–751.
Gu, P., Zhu, Q., & Zhang, C. (2009). A multi-view approach to semi-supervised document classification with incremental naive Bayes. *Computers & Mathematics with Applications, 57*(6), 1030–1036.
Guan, H., Zhou, J., Xiao, B., Guo, M., & Yang, T. (2013). Fast dimension reduction for document classification based on imprecise spectrum analysis. *Information Sciences, 222*, 147–162. doi:10.1016/j.ins.2012.07.032.
Gui, Y., Gao, Z., Li, R., & Yang, X. (2012). Hierarchical text classification for news articles based-on named entities. In *Proceedings of the 2012 international conference on advanced data mining and applications* (pp. 318–329). Springer.
Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.
Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to spam filtering. *Expert Systems with Applications, 36*(7), 10206–10222.
Haddoud, M., Mokhtari, A., Lecroq, T., & Abdeddaïm, S. (2016). Combining supervised term-weighting metrics for SVM text classification with extended term representation. *Knowledge and Information Systems, 49*(3), 909–931. doi:10.1007/s10115-016-0924-1.
Hajmohammadi, M. S., Ibrahim, R., & Selamat, A. (2014). Cross-lingual sentiment classification using multiple source languages in multi-view semi-supervised learning. *Engineering Applications of Artificial Intelligence, 36*, 195–203. doi:10.1016/j.engappai.2014.07.020.
Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. MIT Press.
Herrera, F., Ventura, S., Bello, R., Cornelis, C., Zafra, A., Tarragó, D. S., et al. (2016). *Multiple instance learning – Foundations and algorithms*. Springer. doi:10.1007/978-3-319-47759-6.
Hu, R., Mac Namee, B., & Delany, S. J. (2016). Active learning for text classification with reusability. *Expert Systems with Applications, 45*, 438–449.
Ibáñez, A., Bielza, C., & Larrañaga, P. (2014). Cost-sensitive selective naive Bayes classifiers for predicting the increase of the h-index for scientific journals. *Neurocomputing, 135*, 42–52. doi:10.1016/j.neucom.2013.08.042.
Ittoo, A., Nguyen, L. M., & van den Bosch, A. (2016). Text analytics in industry: Challenges, desiderata and trends. *Computers in Industry, 78*, 96–107. doi:10.1016/j.compind.2015.12.001.
Jalilvand, A., & Salim, N. (2017). Feature unionization: A novel approach for dimension reduction. *Applied Soft Computing, 52*, 1253–1261.
Jiang, L., Li, C., Wang, S., & Zhang, L. (2016). Deep feature weighting for naive Bayes and its application to text classification. *Engineering Applications of Artificial Intelligence, 52*, 26–39.
Jiang, M., Li, N., & Pan, Z. (2016). Multi-label text categorization using L_{21}-norm minimization extreme learning machine. In *Proceedings of the ELM-2015: 90* (pp. 121–133). Springer.
Kamkarhaghighi, M., & Makrehchi, M. (2017). Content tree word embedding for document representation. *Expert Systems with Applications, 90*, 241–249.
Kan, M.-Y., & Thi, H. O. N. (2005). Fast webpage classification using URL features. In *Proceedings of the fourteenth ACM international conference on information and knowledge management, CIKM '05* (pp. 325–326). New York, NY, USA: ACM. doi:10.1145/1099554.1099649.
Kang, M., Ahn, J., & Lee, K. (2018). Opinion mining using ensemble text hidden Markov models for text classification. *Expert Systems with Applications, 94*, 218–227.
Kang, S., Cho, S., & Kang, P. (2015). Multi-class classification via heterogeneous ensemble of one-class classifiers. *Engineering Applications of Artificial Intelligence, 43*, 35–43.
Kang, Y., Haghighi, P. D., & Burstein, F. (2016). Taxofinder: A graph-based approach for taxonomy learning. *IEEE Transactions on Knowledge and Data Engineering, 28*(2), 524–536. doi:10.1109/TKDE.2015.2475759.
Karystianis, G., Thayer, K., Wolfe, M., & Tsafnat, G. (2017). Evaluation of a rule-based method for epidemiological document classification towards the automation of systematic reviews. *Journal of biomedical informatics, 70*, 27–34.
Katakis, I., Tsoumakas, G., & Vlahavas, I. P. (2010). Tracking recurring contexts using ensemble classifiers: An application to email filtering. *Knowledge and Information System, 22*(3), 371–391. doi:10.1007/s10115-009-0206-2.
Ke, W. (2015). Information-theoretic term weighting schemes for document clustering and classification. *International Journal on Digital Libraries, 16*(2), 145–159. doi:10.1007/s00799-014-0121-3.
Kim, H., Howland, P., & Park, H. (2005). Dimension reduction in text classification with support vector machines. *Journal of Machine Learning Research, 6*(Jan), 37–53.
Kim, H. K., & Kim, M. (2016). Model-induced term-weighting schemes for text classification. *Applied Intelligence, 45*(1), 30–43. doi:10.1007/s10489-015-0745-z.
de Knijff, J., Frasincar, F., & Hogenboom, F. (2013). Domain taxonomy learning from text: The subsumption method versus hierarchical clustering. *Data & Knowledge Engineering, 83*, 54–69. doi:10.1016/j.datak.2012.10.002.
Ko, Y., Park, J., & Seo, J. (2004). Improving text categorization using the importance of sentences. *Information Processing and Management, 40*(1), 65–79. doi:10.1016/S0306-4573(02)00056-0.
Kontostathis, A., & Pottenger, W. M. (2006). A framework for understanding latent semantic indexing (LSI) performance. *Information Processing and Management, 42*(1), 56–73. doi:10.1016/j.ipm.2004.11.007.
Kou, G., Lu, Y., Peng, Y., & Shi, Y. (2012). Evaluation of classification algorithms using MCDM and rank correlation. *International Journal of Information Technology and Decision Making, 11*(1), 197–225. doi:10.1142/S0219622012500095.
Král, P. (2014). Named entities as new features for czech document classification. In A. F. Gelbukh (Ed.), *Proceedings of the fifteenth international conference on computational linguistics and intelligent text processing, CICLing 2014, Part II*. In Lecture Notes in Computer Science: *8404* (pp. 417–427). Kathmandu, Nepal: Springer. April 6–12, 2014
Krawczyk, B., Minku, L. L., Gama, J., Stefanowski, J., & Woźniak, M. (2017). Ensemble learning for data stream analysis: A survey. *Information Fusion, 37*, 132–156.
Krohling, R. A., Lourenzutti, R., & Campos, M. (2015). Ranking and comparing evolutionary algorithms with Hellinger–Topsis. *Applied Soft Computing, 37*(C), 217–226. doi:10.1016/j.asoc.2015.08.012.
Kumar, B. S., & Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems, 114*, 128–147.
Labani, M., Moradi, P., Ahmadizar, F., & Jalili, M. (2018). A novel multivariate filter method for feature selection in text classification problems. *Engineering Applications of Artificial Intelligence, 70*, 25–37.
Lee, L. H., Isa, D., Choo, W. O., & Chue, W. Y. (2012). High relevance keyword extraction facility for Bayesian text classification on different domains of varying characteristic. *Expert Systems with Applications, 39*(1), 1147–1155.
Li, C. H., & Park, S. C. (2009). An efficient document classification model using an improved back propagation neural network and singular value decomposition. *Expert Systems with Applications, 36*(2), 3208–3215. doi:10.1016/j.eswa.2008.01.014.
Li, C. H., Yang, J. C., & Park, S. C. (2012). Text categorization algorithms using semantic approaches, corpus-based thesaurus and WordNet. *Expert Systems with Applications, 39*(1), 765–772. doi:10.1016/j.eswa.2011.07.070.
Li, J., Li, J., Fu, X., Masud, M. A., & Huang, J. Z. (2016). Learning distributed word representation with multi-contextual mixed embedding. *Knowledge-Based Systems, 106*, 220–230.
Li, S., Xia, R., Zong, C., & Huang, C.-R. (2009). A framework of feature selection methods for text categorization. In *Proceedings of the joint conference of the forty-seventh annual meeting of the ACL and the fourth international joint conference on natural language processing of the AFNLP, ACL '09: 2* (pp. 692–700). Stroudsburg, PA, USA: Association for Computational Linguistics.
Li, X., Rao, Y., Xie, H., Lau, R. Y. K., Yin, J., & Wang, F. L. (2017). Bootstrapping social emotion classification with semantically rich hybrid neural networks. *IEEE Transactions on Affective Computing, 8*(4), 428–442.
Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems, 69*, 14–23. doi:10.1016/j.knosys.2014.04.022.
Li, Y., Li, T., & Liu, H. (2017). Recent advances in feature selection and its applications. *Knowledge and Information Systems, 53*(3), 551–577.
Li, Y., & Shawe-Taylor, J. (2007). Advanced learning algorithms for cross-language patent retrieval and classification. *Information Processing and Management, 43*(5), 1183–1199. doi:10.1016/j.ipm.2006.11.005. Patent Processing
Lim, C. S., Lee, K. J., & Kim, G. (2005). Multiple sets of features for automatic genre classification of web documents. *Information Processing and Management, 41*(5), 1263–1276. doi:10.1016/j.ipm.2004.06.004.
Lin, S. (2009). A document classification and retrieval system for R&D in semiconductor industry – A hybrid approach. *Expert Systems with Applications, 36*(3), 4753–4764. doi:10.1016/j.eswa.2008.06.024.
Liu, B. (2006). *Web data mining: Exploring hyperlinks, contents, and usage data (data–centric systems and applications)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc..
Liu, B., Xiao, Y., & Hao, Z. (2018). A selective multiple instance transfer learning method for text categorization problems. *Knowledge-Based Systems, 141*, 178–187.
Liu, J. N. K., He, Y., Lim, E. H. Y., & Wang, X. (2014). Domain ontology graph model and its application in Chinese text classification. *Neural Computing and Applications, 24*(3–4), 779–798. doi:10.1007/s00521-012-1272-z.
Liu, Y., Bi, J.-W., & Fan, Z.-P. (2017). Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms. *Expert Systems with Applications, 80*, 323–339.
Liu, Y., Wang, Y., Feng, L., & Zhu, X. (2016). Term frequency combined hybrid feature selection method for spam filtering. *Pattern Analysis and Applications, 19*(2), 369–383. doi:10.1007/s10044-014-0408-4.
Lochter, J. V., Zanetti, R. F., Reller, D., & Almeida, T. A. (2016). Short text opinion detection using ensemble of classifiers and semantic indexing. *Expert Systems with Applications, 62*, 243–249.
Luo, Q., Chen, E., & Xiong, H. (2011). A semantic term weighting scheme for text categorization. *Expert Systems with Applications, 38*(10), 12708–12716. doi:10.1016/j.eswa.2011.04.058.
van der Maaten, L. (2014). Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research, 15*, 3221–3245.
van der Maaten, L., & Hinton, G. E. (2008). Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research, 9*, 2579–2605.
Magerman, T., Looy, B. V., & Song, X. (2010). Exploring the feasibility and accuracy of latent semantic analysis based text mining techniques to detect similarity be-

tween patent documents and scientific publications. *Scientometrics, 82*(2), 289–306. doi:10.1007/s11192-009-0046-6.
Manek, A. S., Shenoy, P. D., Mohan, M. C., & Venugopal, K. (2017). Aspect term extraction for sentiment analysis in large movie reviews using Gini index feature selection method and SVM classifier. *World Wide Web, 20*(2), 135–154.
Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. New York, NY, USA: Cambridge University Press.
Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. New York: Academic Press London.
Matsubara, E. T., Monard, M. C., & Batista, G. E. A. P. A. (2005). Multi-view semi–supervised learning: An approach to obtain different views from text datasets. In K. Nakamatsu, & J. M. Abe (Eds.), *Proceedings of the advances in logic based intelligent systems – selected papers of LAPTEC 2005*. In *Frontiers in Artificial Intelligence and Applications: 132* (pp. 97–104). Himeji, Japan: IOS Press. April 2–4, 2005
Meng, J., Lin, H., & Yu, Y. (2011). A two-stage feature selection method for text categorization. *Computers & Mathematics with Applications, 62*(7), 2793–2800. doi:10.1016/j.camwa.2011.07.045.
Mihalcea, R., & Radev, D. (2011). *Graph-based natural language processing and information retrieval*. Cambridge University Press.
Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic regularities in continuous space word representations.. In *Proceedings of the 2013 conference of the north American chapter of the association for computational linguistics: Human language technologies (NAACL-HLT-2013): 13* (pp. 746–751).
Mostafa, J., & Lam, W. (2000). Automatic classification using supervised learning in a medical document filtering application. *Information Processing and Management, 36*(3), 415–444. doi:10.1016/S0306-4573(99)00033-3.
Muslea, I. (2010). Semi-supervised text processing. In C. Sammut, & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 897–901). Boston, MA: Springer US. doi:10.1007/978-0-387-30164-8_750.
Nanculef, R., Flaounas, I., & Cristianini, N. (2014). Efficient classification of multi-labeled text streams by clashing. *Expert Systems with Applications, 41*(11), 5431–5450.
Nguyen, H.-L., Woon, Y.-K., & Ng, W.-K. (2015). A survey on data stream clustering and classification. *Knowledge and Information Systems, 45*(3), 535–569.
Nunzio, G. M. D. (2014). A new decision to take for cost-sensitive Naïve Bayes classifiers. *Information Processing and Management, 50*(5), 653–674. doi:10.1016/j.ipm.2014.04.008.
Ogura, H., Amano, H., & Kondo, M. (2011). Comparison of metrics for feature selection in imbalanced text classification. *Expert Systems with Applications, 38*(5), 4978–4989. doi:10.1016/j.eswa.2010.09.153.
Olvera-López, J. A., Carrasco-Ochoa, J. A., Trinidad, J. F. M., & Kittler, J. (2010). A review of instance selection methods. *Artificial Intelligence Review, 34*(2), 133–143. doi:10.1007/s10462-010-9165-y.
Onan, A., Korukoglu, S., & Bulut, H. (2016a). Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications, 57*, 232–247. https://doi.org/10.1016/j.eswa.2016.03.045.
Onan, A., Korukoğlu, S., & Bulut, H. (2016b). A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. *Expert Systems with Applications, 62*, 1–16.
Pang, G., & Jiang, S. (2013). A generalized cluster centroid based classifier for text categorization. *Information Processing and Management, 49*(2), 576–586. doi:10.1016/j.ipm.2012.10.003.
Pang, G., Jin, H., & Jiang, S. (2015). CenK: A scalable and effective text classifier. *Data & Knowledge Engineering, 29*(3), 593–625.
Parlak, B., & Uysal, A. K. (2015). Classification of medical documents according to diseases. In *Proceedings of the twenty-third signal processing and communications applications conference (SIU)* (pp. 1635–1638). doi:10.1109/SIU.2015.7130164.
Parmezan, A. R. S., Lee, H. D., & Wu, F. C. (2017). Metalearning for choosing feature selection algorithms in data mining: Proposal of a new framework. *Expert Systems with Applications, 75*, 1–24. doi:10.1016/j.eswa.2017.01.013.
Pavlinek, M., & Podgorelec, V. (2017). Text classification method based on self-training and LDA topic models. *Expert Systems with Applications, 80*, 83–93. doi:10.1016/j.eswa.2017.03.020.
Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).
Pereira, R. B., Plastino, A., Zadrozny, B., & Merschmann, L. H. C. (2018). Categorizing feature selection methods for multi-label classification. *Artificial Intelligence Review, 49*(1), 57–78. doi:10.1007/s10462-016-9516-4.
Perikos, I., & Hatzilygeroudis, I. (2016). Recognizing emotions in text using ensemble of classifiers. *Engineering Applications of Artificial Intelligence, 51*, 191–201. Mining the Humanities: Technologies and Applications
Pi, N. S., Martí, L., & Garcia, A. C. B. (2016). Improving ontology-based text classification: An occupational health and security application. *Journal of Applied Logic, 17*, 48–58. doi:10.1016/j.jal.2015.09.008.
Pillai, I., Fumera, G., & Roli, F. (2017). Designing multi-label classifiers that maximize F -measures: State of the art. *Pattern Recognition, 61*, 394–404.
Pinheiro, R. H., Cavalcanti, G. D., & Tsang, R. (2017). Combining dissimilarity spaces for text categorization. *Information Sciences, 406*, 87–101.
Pinheiro, R. H. W., Cavalcanti, G. D. C., Correa, R. F., & Ren, T. I. (2012). A global-ranking local feature selection method for text categorization. *Expert Systems with Applications, 39*(17), 12851–12857. doi:10.1016/j.eswa.2012.05.008.
Pinheiro, R. H. W., Cavalcanti, G. D. C., & Ren, T. I. (2015). Data-driven global-ranking local feature selection methods for text categorization. *Expert Systems with Applications, 42*(4), 1941–1949. doi:10.1016/j.eswa.2014.10.011.
Qi, X., & Davison, B. D. (2009). Web page classification: Features and algorithms. *ACM Computing Surveys, 41*(2), 12:1–12:31. doi:10.1145/1459352.1459357.
Qin, Z., Cong, Y., & Wan, T. (2016). Topic modeling of Chinese language beyond a bag-of-words. *Computer Speech & Language, 40*, 60–78.
Rao, Y., Li, Q., Wu, Q., Xie, H., Wang, F. L., & Wang, T. (2017). A multi-relational term scheme for first story detection. *Neurocomputing, 254*, 42–52.
Ravi, K., & Ravi, V. (2017). A novel automatic satire and irony detection using ensembled feature selection and data mining. *Knowledge-Based Systems, 120*, 15–33.
Ray, S., Scott, S., & Blockeel, H. (2010). Multi-instance learning. *Encyclopedia of Machine Learning* (pp. 701–710). doi:10.1007/978-0-387-30164-8_569.
Rehman, A., Javed, K., & Babri, H. A. (2017). Feature selection based on a normalized difference measure for text classification. *Information Processing and Management, 53*(2), 473–489.
Ren, F., & Sohrab, M. G. (2013). Class-indexing-based term weighting for automatic text classification. *Information Sciences, 236*, 109–125. doi:10.1016/j.ins.2013.02.029.
Ripley, B. D., & Hjort, N. L. (1995). *Pattern recognition and neural networks* (1st). New York, NY, USA: Cambridge University Press.
Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., & Gatford, M. (1995). Okapi at TREC-3. *Overview of the Third Text REtrieval Conference (TREC3)* (pp. 109–126). Gaithersburg, MD: NIST.
Rossi, R. G., de Andrade Lopes, A., & Rezende, S. O. (2017). Using bipartite heterogeneous networks to speed up inductive semi-supervised learning and improve automatic text categorization. *Knowledge-Based Systems, 132*, 94–118.
Ruiz, M. E., & Srinivasan, P. (2002). Hierarchical text categorization using neural networks. *Information Retrieval Journal, 5*(1), 87–118.
Sabbah, T., Selamat, A., Selamat, M. H., Al-Anzi, F. S., Viedma, E. H., Krejcar, O., et al. (2017). Modified frequency-based term weighting schemes for text classification. *Applied Soft Computing, 58*, 193–206.
Sabbah, T., Selamat, A., Selamat, M. H., Ibrahim, R., & Fujita, H. (2016). Hybridized term-weighting method for dark web classification. *Neurocomputing, 173*, 1908–1926. doi:10.1016/j.neucom.2015.09.063.
Saha, S., & Ekbal, A. (2013). Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition. *Data & Knowledge Engineering, 85*, 15–39. doi:10.1016/j.datak.2012.06.003.
Sajgalik, M., Barla, M., & Bielikova, M. (2017). Searching for discriminative words in multidimensional continuous feature space. *Computer Speech & Language*, In Press.
Saleh, A. I., Abulwafa, A. E., & Al Rahmawy, M. F. (2017). A web page distillation strategy for efficient focused crawling based on optimized Naïve Bayes (ONB) classifier. *Applied Soft Computing, 53*, 181–204.
Saleh, A. I., Al Rahmawy, M. F., & Abulwafa, A. E. (2017). A semantic based web page classification strategy using multi-layered domain ontology. *World Wide Web, 20*(5), 939–993.
Saloot, M. A., Idris, N., Mahmud, R., Jaafar, S., Thorleuchter, D., & Gani, A. (2016). Hadith data mining and classification: A comparative analysis. *Artificial Intelligence Review, 46*(1), 113–128.
Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers, 18*(5), 401–409. doi:10.1109/T-C.1969.222678.
(2010a). Cross-validation. In C. Sammut, & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 249–249)). Boston, MA: Springer US. doi:10.1007/978-0-387-30164-8_190.
(2010b). Roc curve. In C. Sammut, & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 875–875)). Boston, MA: Springer US. doi:10.1007/978-0-387-30164-8_735.
(2010c). Stacked generalization. In C. Sammut, & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 912–912)). Boston, MA: Springer US. doi:10.1007/978-0-387-30164-8_778.
(2017). Supervised learning. In C. Sammut, & G. I. Webb (Eds.), *Encyclopedia of Machine Learning and Data Mining* (pp. 1213–1214)). Boston, MA: Springer US. doi:10.1007/978-1-4899-7687-1_803.
Santafe, G., Inza, I., & Lozano, J. A. (2015). Dealing with the evaluation of supervised classification algorithms. *Artificial Intelligence Review, 44*(4), 467–508. doi:10.1007/s10462-015-9433-y.
Schenker, A., Bunke, H., Last, M., & Kandel, A. (2005). Graph-theoretic techniques for web content mining. Series in machine perception and artificial intelligence, *Vol.* 62.
Schölkopf, B., Smola, A., & Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation, 10*(5), 1299–1319. doi:10.1162/089976698300017467.
Schürmann, J. (1996). *Pattern classification – A unified view of statistical and neural approaches.*. Wiley.
Seijo-Pardo, B., Porto-Díaz, I., Bolón-Canedo, V., & Alonso-Betanzos, A. (2017). Ensemble feature selection: Homogeneous and heterogeneous approaches. *Knowledge-Based Systems, 118*, 124–139. doi:10.1016/j.knosys.2016.11.017.
Sharef, N. M., & Martin, T. (2015). Evolving fuzzy grammar for crime texts categorization. *Applied Soft Computing, 28*, 175–187.
Sheikhpour, R., Sarram, M. A., Gharaghani, S., & Chahooki, M. A. Z. (2017). A survey on semi-supervised feature selection methods. *Pattern Recognition, 64*, 141–158. doi:10.1016/j.patcog.2016.11.003.
Shen, L., Chen, H., Yu, Z., Kang, W., Zhang, B., Li, H., et al. (2016). Evolving support vector machines using fruit fly optimization for medical data classification. *Knowledge-Based Systems, 96*, 61–75.
Shi, L., Ma, X., Xi, L., Duan, Q., & Zhao, J. (2011). Rough set and ensemble learning based semi-supervised algorithm for text classification. *Expert Systems with Applications, 38*(5), 6300–6306.

Silva, R. M., Almeida, T. A., & Yamakami, A. (2017). MDLText: An efficient and lightweight text classifier. *Knowledge-Based Systems, 118*, 152–164.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management, 45*(4), 427–437.

Sun, B., Feng, J., & Saenko, K. (2016). Return of frustratingly easy domain adaptation.. In *Proceedings of the thirtieth AAAI conference on artificial intelligence: 6* (p. 8).

Sun, S. (2013). A survey of multi-view machine learning. *Neural Computing and Applications, 23*(7–8), 2031–2038.

Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information Fusion, 36*, 10–25.

Tan, B., Zhong, E., Xiang, E. W., & Yang, Q. (2014). Multi-transfer: Transfer learning with multiple views and multiple sources. *Statistical Analysis and Data Mining, 7*(4), 282–293. doi:10.1002/sam.11226.

Tang, B., He, H., Baggenstoss, P. M., & Kay, S. (2016). A Bayesian classification approach using class-specific features for text categorization. *IEEE Transactions on Knowledge and Data Engineering, 28*(6), 1602–1606. doi:10.1109/TKDE.2016.2522427.

Tang, B., Kay, S., & He, H. (2016). Toward optimal feature selection in naive Bayes for text categorization. *IEEE Transactions on Knowledge and Data Engineering, 28*(9), 2508–2521. doi:10.1109/TKDE.2016.2563436.

Thorleuchter, D., & den Poel, D. V. (2013). Technology classification with latent semantic indexing. *Expert Systems with Applications, 40*(5), 1786–1795. doi:10.1016/j.eswa.2012.09.023.

Ting, K. M., & Witten, I. H. (1997). Stacked generalization: when does it work? In *Proceedings of the 1997 international joint conference on artificial intelligence* (pp. 866–871). Morgan Kaufmann.

Ting, K. M., & Witten, I. H. (1999). Issues in stacked generalization. *Journal of Artificial Intelligence Research, 10*, 271–289.

Tommasel, A., & Godoy, D. (2018). Short-text feature construction and selection in social media data: a survey. *Artificial Intelligence Review, 49*(3), 301–338. doi:10.1007/s10462-016-9528-0.

Trappey, A. J. C., Hsu, F., Trappey, C. V., & Lin, C. (2006). Development of a patent document classification and search platform using a back-propagation network. *Expert Systems with Applications, 31*(4), 755–765. doi:10.1016/j.eswa.2006.01.013.

Tripathy, A., Anand, A., & Rath, S. K. (2017). Document-level sentiment classification using hybrid machine learning approach. *Knowledge and Information Systems, 53*(3), 805–831. doi:10.1007/s10115-017-1055-z.

Tsai, C., & Chang, C. (2013). SVOIS: Support vector oriented instance selection for text classification. *Information Sciences, 38*(8), 1070–1083. doi:10.1016/j.is.2013.05.001.

Tutkan, M., Ganiz, M. C., & Akyokuş, S. (2016). Helmholtz principle based supervised and unsupervised feature selection methods for text mining. *Information Processing and Management, 52*(5), 885–910.

Uguz, H. (2011). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowledge-Based Systems, 24*(7), 1024–1032. doi:10.1016/j.knosys.2011.04.014.

Uysal, A. K. (2016). An improved global feature selection scheme for text classification. *Expert Systems with Applications, 43*, 82–92.

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

Van Linh, N., Anh, N. K., Than, K., & Dang, C. N. (2017). An effective and interpretable method for document classification. *Knowledge and Information Systems, 50*(3), 763–793.

Vanderlooy, S., & Hüllermeier, E. (2008). A critical analysis of variants of the AUC. *Machine Learning, 72*(3), 247–262. doi:10.1007/s10994-008-5070-x.

Vergara, J. R., & Estévez, P. A. (2014). A review of feature selection methods based on mutual information. *Neural Computing and Applications, 24*(1), 175–186. doi:10.1007/s00521-013-1368-0.

Vilalta, R., Giraud-Carrier, C., Brazdil, P., & Soares, C. (2010). Inductive transfer. In C. Sammut, & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 545–548). Boston, MA: Springer US. doi:10.1007/978-0-387-30164-8_401.

Vinodhini, G., & Chandrasekaran, R. (2016). A comparative performance evaluation of neural network based approach for sentiment classification of online reviews. *Journal of King Saud University-Computer and Information Sciences, 28*(1), 2–12.

Vlachos, M. (2010). Dimensionality reduction. In C. Sammut, & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 274–279)). Boston, MA: Springer US. doi:10.1007/978-0-387-30164-8_216.

Wang, D., Wu, J., Zhang, H., Xu, K., & Lin, M. (2013). Towards enhancing centroid classifier for text classificationa border-instance approach. *Neurocomputing, 101*, 299–308.

Wang, P., Hu, J., Zeng, H., & Chen, Z. (2009). Using Wikipedia knowledge to improve text classification. *Knowledge and Information Systems, 19*(3), 265–281. doi:10.1007/s10115-008-0152-4.

Wang, P., Xu, B., Xu, J., Tian, G., Liu, C.-L., & Hao, H. (2016). Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. *Neurocomputing, 174*, 806–814.

Wang, S., Li, D., Song, X., Wei, Y., & Li, H. (2011). A feature selection method based on improved Fisher's discriminant ratio for text sentiment classification. *Expert Systems with Applications, 38*(7), 8696–8702. doi:10.1016/j.eswa.2011.01.077.

Wang, W., & Yu, B. (2009). Text categorization based on combination of modified back propagation neural network and latent semantic analysis. *Neural Computing and Applications, 18*(8), 875–881. doi:10.1007/s00521-008-0193-3.

Wang, Y., Wang, J., Liao, H., & Chen, H. (2017). An efficient semi-supervised representatives feature selection algorithm based on information theory. *Pattern Recognition, 61*, 511–523. doi:10.1016/j.patcog.2016.08.011.

Wei, C.-P., Lin, Y.-T., & Yang, C. C. (2011). Cross-lingual text categorization: Conquering language boundaries in globalized environments. *Information Processing and Management, 47*(5), 786–804.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data, 3*(1), 9.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks, 5*, 241–259.

Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition, 48*(9), 2839–2846. doi:10.1016/j.patcog.2015.03.009.

Woźniak, M., Graña, M., & Corchado, E. (2014). A survey of multiple classifier systems as hybrid systems. *Information Fusion, 16*, 3–17.

Wu, Z., Zhu, H., Li, G., Cui, Z., Huang, H., Li, J., et al. (2017). An efficient Wikipedia semantic matching approach to text document classification. *Information Sciences, 393*, 15–28. doi:10.1016/j.ins.2017.02.009.

Xia, R., Xu, F., Yu, J., Qi, Y., & Cambria, E. (2016). Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis. *Information Processing and Management, 52*(1), 36–45.

Xia, R., Zong, C., & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences, 181*(6), 1138–1152.

Xiao, Y., Liu, B., Yin, J., & Hao, Z. (2017). A multiple-instance stream learning framework for adaptive document categorization. *Knowledge-Based Systems, 120*, 198–210. doi:10.1016/j.knosys.2017.01.001.

Xie, F., Wu, X., & Zhu, X. (2017). Efficient sequential pattern mining with wildcards for keyphrase extraction. *Knowledge-Based Systems, 115*, 27–39. doi:10.1016/j.knosys.2016.10.011.

Xu, C., Tao, D., & Xu, C. (2013). A survey on multi-view learning. arXiv preprint: 1304.5634.

Yan, K., Li, Z., & Zhang, C. (2016). A new multi-instance multi-label learning approach for image and text classification. *Multimedia Tools and Applications, 75*(13), 7875–7890.

Yan, S., Zhu, X., Liu, G., & Wu, J. (2017). Sparse multiple instance learning as document classification. *Multimedia Tools and Applications, 76*(3), 4553–4570.

Yang, B., Zhang, Y., & Li, X. (2011). Classifying text streams by keywords using classifier ensemble. *Data & Knowledge Engineering, 70*(9), 775–793.

Yin, C., & Xi, J. (2017). Maximum entropy model for mobile text classification in cloud computing using improved information gain algorithm. *Multimedia Tools and Applications, 76*(16), 16875–16891. doi:10.1007/s11042-016-3545-5.

Yoon, Y., & Lee, G. G. (2007). Efficient implementation of associative classifiers for document classification. *Information Processing and Management, 43*(2), 393–405. doi:10.1016/j.ipm.2006.07.012.

Yu, K., Wu, X., Ding, W., & Pei, J. (2016). Scalable and accurate online feature selection for big data. *ACM Transactions on Knowledge Discovery from Data, 11*(2), 16:1–16:39. doi:10.1145/2976744.

Zelaia, A., Alegria, I., Arregi, O., & Sierra, B. (2011). A multiclass/multilabel document categorization system: Combining multiple classifiers in a reduced dimension. *Applied Soft Computing, 11*(8), 4981–4990.

Zhang, H., & Zhong, G. (2016). Improving short text classification by learning vector representations of both words and hidden topics. *Knowledge-Based Systems, 102*, 76–86. doi:10.1016/j.knosys.2016.03.027.

Zhang, J., Chen, L., & Guo, G. (2013). Projected-prototype based classifier for text categorization. *Knowledge-Based Systems, 49*, 179–189.

Zhang, L., Jiang, L., Li, C., & Kong, G. (2016). Two feature weighting approaches for naive Bayes text classifiers. *Knowledge-Based Systems, 100*, 137–144.

Zhang, Y., Chu, G., Li, P.-P., Hu, X., & Wu, X. (2017). Three-layer concept drifting detection in text data streams. *Neurocomputing, 260*, 393–403. doi:10.1016/j.neucom.2017.04.047.

Zhao, J., Xie, X., Xu, X., & Sun, S. (2017). Multi-view learning overview: Recent progress and new challenges. *Information Fusion, 38*, 43–54.

Zhou, G., Zhu, Z., He, T., & Hu, X. T. (2016). Cross-lingual sentiment classification with stacked autoencoders. *Knowledge and Information Systems, 47*(1), 27–44. doi:10.1007/s10115-015-0849-0.

Zhou, Y., Jin, R., & Hoi, S. C. H. (2010). Exclusive lasso for multi-task feature selection. In Y. W. Teh, & D. M. Titterington (Eds.), *Proceedings of the thirteenth international conference on artificial intelligence and statistics, AISTATS 2010: 9* (pp. 988–995). Chia Laguna Resort, Sardinia, Italy: JMLR.org. May 13–15, 2010

Zhu, X. (2010). Semi-supervised learning. In C. Sammut, & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 892–897). Boston, MA: Springer US. doi:10.1007/978-0-387-30164-8_749.

Zuo, Y., Zhao, J., & Xu, K. (2016). Word network topic model: A simple but general solution for short and imbalanced texts. *Knowledge and Information Systems, 48*(2), 379–398.