Automatic Question Classifiers: a Systematic Review

Valtemir A. Silva, Ig I.B. S. Pinto, and José C. Maldonado

Abstract— Question classification is a key point in many applications, such as Question Answering (QA), Information Retrieval (IR) and E-learning systems. This paper aims to carry out a systematic review of the literature on automatic question classifiers and the technology directly involved. Automatic classifiers are responsible for labeling a certain evaluation item using a type of categorization as a selection criterion. The analysis of 80 primary studies previously selected revealed that SVM is the main algorithm of the Machine Learning used, while BOW and TF-IDF are the main techniques for feature extraction and selection, respectively. According to the analysis, the taxonomies proposed by Li and Roth and Bloom were the most used ones for the classification criteria, and Accuracy/Precision/Recall/F1-score were proven to be the most used metrics. In the future, the objective is to perform a meta-analysis with the studies that authorize the availability of their data.

Index Terms—Question classification, machine learning, feature selection, feature extraction.

1 Introduction

The large amount of digital information available on the Internet, especially in the form of text, transforms the knowledge organization, analysis and extraction into essential activities, both in the academic universe and in the job market. For this reason, the automatic classification of texts has been gaining more and more prominence in these tasks [1], [2].

Within the universe of text or document classification, there is a more restricted subgroup called question classification that basically corresponds to the association of a label according to a pre-determined criterion. The problem becomes complex once the amount of information available in a question is much smaller when compared to the texts in documents in general [3].

The main applications of automatic question classification are [4]:

1. *Question Answering (QA):* primary application of question classification based on questions formulated in natural language aimed to retrieve a set of associated documents and find the most compatible answers. Example: Yahoo! answers.
2. *Information Retrieval (IR):* similar to QA, it focuses on the document retrieval. Example: Google's search engine.
3. *E-learning:* retrieval of questions – normally multiple-choice – divided into categories – competence or degree of difficulty – for cognitive evaluation tests. Example: classifiers via Bloom's taxonomy.
4. *Specific languages:* document retrieval and conversion evolving specific language characteristics not present in the English language. Example: Chinese classifiers.

Specifically considering the educational environment, the automatic generation of evaluation tests has immediate and practical application in e-learning systems since it enables the customization of teaching by searching for questions appropriate to a certain learning profile [5]. These adaptive teaching systems make use of question banks of various formats and use them to apply diagnostic tests on different skills and competencies on a continuous basis [6]. To use these questions in tests, it is vital to classify them in terms of their skill area and degree of difficulty in order to appropriately recommend them to each student according to their performance [7]. Nevertheless, creating a bank of classified items is a complex task due to the need of categorizing a large number of questions in a representative variety of skills and competences, among other reasons. On the other hand, there is a vast number of nonclassified questions in digital didactic materials and on the web that could be used for making educational diagnosis were they related to the skills and competences they evaluate.

Considering such limitations, this paper aims to describe the current scenario evolving techniques and algorithms for question classification. To do so, a systematic review (SR) was conducted to identify, evaluate, interpret and synthesize the primary studies available to establish the state of the art in this area.

Some works have already been done considering question classification, although not exclusively focusing on such subject and using a smaller number of studies as initial selection [4], [8], [9], [10]. The idea of this work is to present a detailed view of the question classification area considering the most used techniques, criteria and indicators. In this sense, this paper shows the results of a SR with studies published between January 2012 and July 2017, following a protocol defined according to the guidelines proposed by [11], [12].

The work is organized as follows: in Chapter 2, the main techniques associated with question classification are

- *V.A. Silva and J.C. Maldonado ~~are with~~work at the Institute of Mathematics and Computer Science University of São Paulo, 13566-590, São Carlos, SP, Brazil. E-mail: valtemir.alencar@usp.br, jcmaldon@icmc.usp.br.*
- *I.I. Bittencourt ~~is with~~works at the Computing Institute, Federal University of Alagoas, BR 104 Norte km 97, 57072-970 Maceió, AL, Brazil. E-mail: ig.ibert@ic.ufal.br.*

described, including algorithms, taxonomies and result metrics; in Chapter 3, the methodology adopted to execute the SR activities are presented; in Chapter 4, the results of quality evaluation on the primary studies are described, an overview of the selected studies is provided and an analysis of the extracted data is performed; finally, Chapter 5 discusses the scope of the work, lists the threats to its validity and the challenges encountered, points out possible future work, and presents the final conclusions.

2 Automatic Question Classification

Although question classification works in a similar way to text classification, the difficulty in obtaining reasonable precision indexes is huge [3]. This happens as a result of the small amount of information in the questions when compared to the information present in text documents. The main idea is to effectively associate labels to questions according to the intended criterion, which can be related to an assignment of a degree of difficulty (e.g. easy, hard), types of expected answers (with subjects such as politics, sports, health etc.) or even a specific school topic (e.g. Geometry/Mathematics or Syntax Analysis/Portuguese), among others.

One of the first outstanding works developed considering question classification that is still used as reference is the study described by [3], developed at the University of Illinois Urbana-Champaign (UIUC). It is a SNoW hierarchical classifier with a base of 6,000 English questions for classification in 6 main categories and 50 specific ones (Fig. 1). The classifier presented an accuracy of 84.2% and its results and database are still a reference in the area, including the majority of studies selected in this SR.

Class	Class	Class	Class
ABBREV	letter	individual	**NUMERIC**
exp	other	title	code
abb	plant	description	count
ENTITY	product	**LOCATION**	date
animal	religion	city	distance
body	sport	country	money
color	substance	mountain	order
creative	symbol	other	other
currency	technique	state	period
dismed	term	**DESCRIPTION**	percent
event	vehicle	definition	speed
food	word	description	temp
instrument	**HUMAN**	manner	size
lang	group	reason	weight

Fig. 1. Categories defined by Li and Roth [4].

Another work that is a reference in the area is the one done by [13], who used a compact set of features through head words and hypernyms – such terms will be later explained – with an accuracy of 89.2% using the Linear SVM algorithm and 89% with the Maximum Entropy (ME) algorithm.

[14] describes several ways of dividing types of classifiers. Taking into consideration the number of classification categories, they can be divided into binary (2 classes, a single association with the question), single-label (several classes, a single association with the question) e multi-label (several classes, several associations). On the other hand, considering the relationship between the categories, the classifiers can be divided into flat (no relationship) and hierarchical (categories and sub-categories). As to classification decision, they can be grouped into hard (true/false to define if a question belongs to a category) and soft (numeric indicator that measures the degree of reliability of question classification within a certain category). According to [8], [15], [16], the main way of categorizing question classifiers is in relation to their class association strategy, consisting of 3 types:

1. *Rule-based:* nonstatistical approach by means of formulas based on question structure to predict classes with a high degree of accuracy that decreases with a large number of questions.
2. *Machine Learning (ML):* search of patterns through statistical analysis of the structure of a set of questions. Its degree of correction in classification tends to improve by increasing the amount of data. It can be supervised (model generated from a set of labeled data used as training), unsupervised (without set of data labeled for training) and semi-supervised (model generated from an initial set of labeled and unlabeled data).
3. *Hybrid:* a combination of rules and ML algorithm.

Based on the work of [10], Fig. 2 shows the different stages during question classification. The idea is to transform the initial text of the question into a set of relevant features that directly influence on the classifier performance.

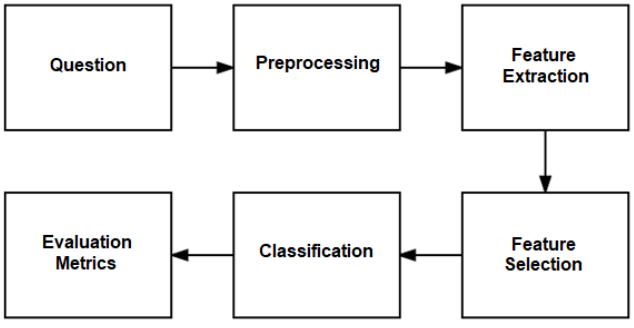

Fig. 2. Stages of the classification process [10].

2.1 Preprocessing

According to [14], the preprocessing phase consists of converting the initial text of the question into a well-defined set of features. This includes removing all irrelevant words such as pronouns, prepositions, punctuation, as well as unifying similar words. Among the main possible preprocessing operations, we can highlight [14], [17]:

1. *Tokenization:* division of the text into elementary fragments called tokens, which are separated by a specific character. In the case of questions, the tokens are the words, and the specific character is the space.
2. *Removal of Stop Words:* withdrawal of common words such as pronouns, prepositions and articles

(a)

Feature Space	Features
Unigram	{(Who, 1) (was, 1) (elected, 1) (president, 1) (of, 1) (South, 1) (Africa, 1) (in, 1) (1994, 1) (?, 1)}
Bigram	{(Who-was, 1), (was-elected, 1), (elected-president, 1), (president-of, 1), (of-South, 1), (South-Africa, 1), (Africa-in, 1), (in-1994, 1), (1994-?, 1)}
Trigram	{(Who-was-elected, 1), (was-elected-president, 1), …, (in-1994-?, 1)}
Wh-Word	{(Who, 1)}
Word-Shapes	{(lowercase, 5) (mix, 3) (digit, 1) (other, 1)}

(b)

Question	Category
What city has the zip code of 35824 ?	LOC:city
Who developed the vaccination against polio ?	HUM:ind
Who invented the slinky ?	HUM:ind
George Bush purchased a small interest in which baseball team ?	HUM:gr
When did Idaho become a state ?	NUM:date
What river flows between Fargo, North Dakota and Moorhead, Minnesota ?	LOC:other
What is the oldest city in Spain ?	LOC:city

(c)

"What river flows between Fargo, North Dakota and Moorhead. Minnesota?".

Feature Space	Features
Hypernyms	{(river, 1) (stream, 1) (body-of-water, 1) (thing, 1) (physical-entity, 1) (entity, 1)}
Related Words	{(rel:What, 1) (rel:list.tar, 2) (rel:loca, 2)}
Question Category	{(other, 1)}
Query Expansion	{(river, 1) (stream, 0.6) (body-of-water, 0.36) (thing, 0.22) (physical-entity, 0.13) (entity, 0.08)}

Fig. 3. Features types [18]: (a) Lexical, (b) Syntactic and (c) Semantic.

(all called stop words), which have no relevance to a certain classification. Other irrelevant characters, such as numeric ones and punctuation, are also deleted.

3. *Tagging:* association of morphological-lexical classes with each token, e.g. article, verb and adverb.
4. *Stemming:* replacement of each token with its word stem, e.g. 'writer', 'writing' and 'wrote' with 'write'.
5. *Parsing:* generation of a token structure in a tree format from a set of grammatical rules in order to represent the syntactic structure of the question.

2.2 Feature Extraction

As a result of the preprocessing phase, several types of features can be extracted and, consequently, can directly influence on the performance of the classification process according to the strategy adopted. Usually, they are divided into lexical, syntactic and semantic [15], [17] (see Fig. 3):

1. *Lexical:* basically, it consists of words in a question context. Below are some examples:
   a. *Unigram or Bag-of-words (BOW):* they correspond to each pair (t, f), where t represents a question word, and f is the number of times this word t appears.
   b. *N-grams:* generalization of BOW for N consecutive terms that appear f times in a question. Unigrams is the case for N=1, Bigrams is the case for N=2, and so on.
   c. *Wh-words:* words that imply a question sense (what, who, where etc.).
2. *Syntactic:* features derived from the syntactic structure of the question. Here are some examples:
   a. *POS tags (Part-of-Speech tags):* tagging of words according to their word class or words such as adjectives, nouns, adverbs etc.
   b. *Head words:* keywords in questions obtained through a parsing operation and that contain the main information for the classification process (Fig. 3).
3. *Semantic:* features associated with a certain question classification. The main examples are Hypernyms/Hiponyms, that is, words that represent the semantic concept of generalization/specialization. For instance: (Color)/(Blue, Red, Green) or (Flower)/(Rose, Jasmime and Orchid).

2.3 Feature selection

Feature selection consists of discovering which features are more relevant than others for the classification problem. Reducing the number of features tends to make the classification algorithm faster and more efficient by improving the performance indicators, such as accuracy. The idea is to use algorithms to calculate weights and/or indexes in order to differentiate the most relevant features. Among the most common techniques for feature selection we can highlight:

1. *Binary* [19]: simple technique consisting of assigning the value 1 if the feature appears in a given question, and 0 if there is no occurrence.
2. *DF - Document Frequency* [20]: number of documents (questions) in which a certain feature occurs. A limit value is established and all features with

smaller DF values are discarded.

3. *TF – Term Frequency* [19]: corresponds to the total occurrences of a feature in the same question.
4. *TF-IDF – Inverse Document Frequency* [19]: this technique adjusts the TF value once common terms in many questions usually do not contribute to the classification and appear very frequently. The formula for calculating the TF-IDF for each feature is:

$$TFIDF = \mathrm{TF} * \log(\mathrm{N}/\mathrm{n})$$

Where:
N = Total number of questions
n = number of questions in which the feature appears (DF).

5. *TF-ICF – Inverse Class Frequency* [19]: method similar to TF-IDF, it considers the incidence of a feature in many categories instead of questions. It works as an adjustment factor to the original TF-IDF:

$$TFICF = \mathrm{TF} * [1 + \log(\mathrm{N}/\mathrm{n})] * [1 + \log(\mathrm{C}/\mathrm{c})]$$

Where:
N = Total number of questions
n = number of questions in which the feature appears (DF).
C = Total number of categories
c = number of categories in which the feature appears.

6. *CHI² - Chi-Square* [14]: measure of the degree of dependence between a feature *f* and a category *c*. The higher the value of this statistic, the higher the association between *f* and *c*. to calculate it, it is necessary to take as a basis the contingency table between *f* and *c* (Table 1), where A is the number of times *f* appears in *c*, B is the number of times that *f* appears without *c*, E is the number of times *c* appears without *f*, K is the number of times *f* and *c* do not appear, and N is the total number of documents:

$$CHI^2(\mathrm{f},\mathrm{c}) = \frac{\mathrm{N}(\mathrm{AK} - \mathrm{EB})^2}{(\mathrm{A}+\mathrm{E})(\mathrm{B}+\mathrm{K})(\mathrm{A}+\mathrm{B})(\mathrm{E}+\mathrm{K})}$$

Other techniques widely used are: Information Gain (IG), Odds Ratio (OR) and Mutual Information (MI),

TABLE 1
CONTINGENCY TABLE BETWEEN F AND C

	c	***NOT c***	***Total***
f	A	B	$A+B$
NOT f	E	K	$E+K$
Total	$A+E$	$B+K$	N

which are detailed by [8].

2.4 Algorithms

As described at the beginning of the chapter, question classifiers are practically divided into rule-based algorithms, machine learning algorithms, and a combination between them. The rules are usually hard-coded instructions, that is, formulas defined according to the problem, having no predefined algorithms. Machine Learning has a great variety of algorithms and the following sections highlight the most used ones in question classification.

2.4.1 KNN - K-Nearest Neighbor

According to [10], KNN is an algorithm in which objects are classified through voting of several training examples labeled with their smallest possible distances for each object. The algorithm is known for its ability to recognize patterns, and its major disadvantage is the need of using all features to compute distances, which generates a high computational cost [10], [21].

2.4.2 Decision trees

A decision tree reconstructs training data in a tree format using well-defined conditions in a true-false format [21]. The leaf nodes represent a set of features, and consequently one or more groups or categories. With the generated tree, new data can be confronted to the conditions of each node, resulting in a prediction of classification for them.

The main advantage is the ease of understanding the model through its structure, while the main disadvantage is the fact that training data have many features resulting in performance problems and overfitting [20]. There is a variety of algorithms for decision trees, such as ID3, ID4-hat, ID5, C4.5 and CART.

2.4.3 Naive Bayes

The Naive Bayes algorithm is also a known algorithm in the text classification area and basically consists of the analysis of probabilities from a training database; these probabilities classify questions through the extracted features [10]. Naive Bayes assumes an independence between features, making the learning performance fast and simple. The problem is the loss of performance in case there is a kind of relationship between the features.

2.4.4 SVM – Support Vector Machine

SVM (Support Vector Machine) is a classification method for linear and nonlinear data. It uses nonlinear mapping to transform training data into higher dimensions, and then finds a linear optimal hyperplane for category separation [10].

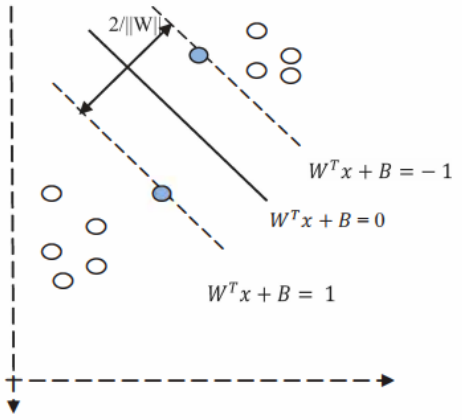

Fig. 4. Stages of the classification process [20].

The linear classifier is defined by the function WT x + B = 0 (Fig. 4), where W is the hyperplane direction and B is its exact position [20].

The items outside the hyperplanes represent two separate categories, and the coordinates belonging to the hyperplanes are known as support vectors. For [20], SVM is robust and has optimal accuracy values, although it is highly complex and requires extensive memory usage for large-scale tasks.

2.4.5 ANN – Artificial Neural Network

ANN is a system composed of processing units called neurons that receive inputs, and each input generates an output. According to [21], the objective is to simulate some of the functionalities of the biological neuron through the interconnection between neurons. Its structure can range from a simple layer of neurons (perceptron) to multiple layers such as MLP (multilayer perceptron), as shown in Fig 5:

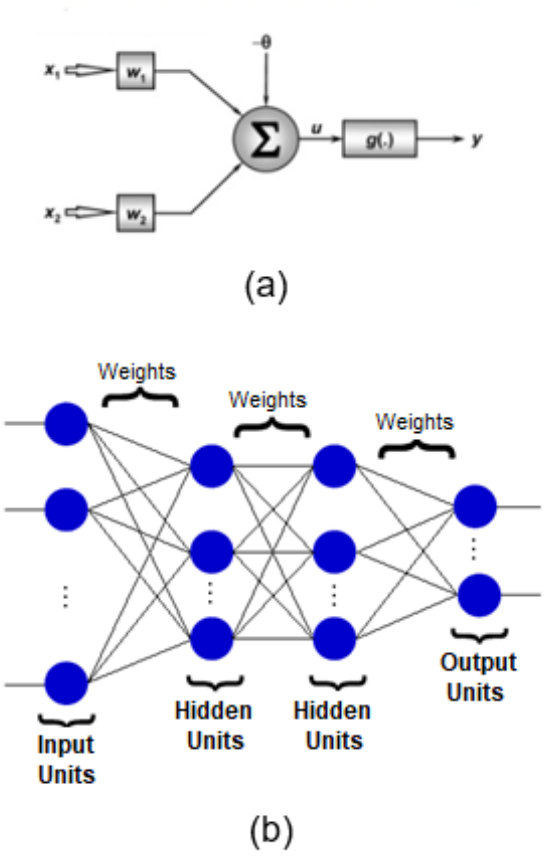

Fig. 5. Neural networks: (a) Perceptron and (b) MLP.

Still according to the authors, one advantage is the good performance with a great number of features, consequently having as a disadvantage the high cost of processing and memory, besides the inherent difficulty in understanding the ANN operation by its users.

2.5 Taxonomies

The process of question classification requires criteria to separate categories. They can be expressed through taxonomies, which detail the common requirements for grouping.

As shown in Fig. 1, the taxonomy proposed by [3] is one of the most used because of its well-defined classes and database available for use.

Another widely referenced taxonomy is the one proposed by Bloom, which consists of a system based on educational objectives to classify questions according to learning and comprehension levels [22]. The taxonomy has three domains: cognitive, affective and psychomotor, although some works on classification of evaluation items usually consider only the cognitive domain. This domain is similar to school evaluations and consists of six hierarchical levels, starting from the simplest one, knowledge, to the most complex one, evaluation. Table 2 describes each level of the cognitive domain illustrating simple examples:

TABLE 2
BLOOM'S TAXONOMY

Category	Example of Questions
Knowledge	Define the concept of inheritance
Comprehension	Explain the structure of a method in a program
Application	Demonstrate the relationship between packages, classes and methods
Analysis	List the advantages of using a container-type class
Synthesis	Create a Java program showing the concept of overload
Evaluation	Justify the concept of inheritance and write a sample source code

2.6 Performance Indicators

The great goal of automatic classifiers is to get as close as possible to the accuracy of label assignment, thereby reducing human intervention. To do so, it is necessary to evaluate the process progress, which can be accomplished through performance indicators.

The indicators are extracted by means of surveying correctly categorized questions and classification errors, which are obtained through the generated contingency matrix. The contingency matrix is generally the basis for most indicators (Table 3), although there are others, such as processing time.

The most known indicators are [8], [14]:

1. *Accuracy:* proportion of correctly classified questions.
   Accuracy = (TP + TN) / TOTAL
   Where:
   TP: True Positives (questions correctly classified in the category)
   TN: True Negatives (questions correctly nonclassified in the category)
   TOTAL = Total number of questions
2. *Precision:* proportion of correctness between the questions predicted for a certain category.
   Precision = TP / (TP + FP)
   Where:
   TP: True Positives (questions correctly classified in the category)

TABLE 3
CONTINGENCY MATRIX FOR THE CLASSIFICATION PROCESS – ADAPTED FROM [23]

		Current Classification			
		Positive Classification	Negative Classification		
Predicted Classif.	Positive Prediction	**TP** (*True Positives*)	**FP** (*False Positives*)	**Precision (P) = TP/ (TP + FP)**	
	Negative Prediction	**FN** (*False Negatives*)	**TN** (*True Negatives*)		
		Recall (R) = TP/ (TP + FN)		**Accuracy = (TP + TN)/ Total**	**F1-score = 2*P*R/ (P + R)**

FP: False Positives (questions incorrectly classified in the category)
TP + FP = Total number of questions classified in a certain category

3. *Recall:* proportion of correctness for the questions belonging to a certain category.
   Recall = TP / (TP + FN)
   Where:
   TP: True Positives (questions correctly classified in the category)
   FN: True Negatives (questions incorrectly nonclassified in the category)
   TP + TN = Total number of questions belonging to a certain category
4. *F1-score (F1-measure):* Harmonic mean of Precision and Recall.
   F1 = (2 * Precision * Recall) / (Precision + Recall)

3 METHOD

Using an SR as a research method means to identify, evaluate and interpret a search for information associated with a research question, area or phenomenon to generate evidence that may support possible conclusions [24]. The context in which the SR is performed is the automatic question classification, also called evaluation items, as introduced in Chapter 2.

The SR protocol guidelines and template were based on the work of [11], [12], [24], [25] and and include the activities of planning, execution and result analysis.

The SR planning is responsible for identifying whether there is a need for a systematic review on the chosen topic and, if so, which strategy should be used to search for primary studies through a protocol definition [12]. Such protocol includes the methods for defining the research question, search terms and sources, selection and exclusion criteria as well as data extraction and synthesis. The protocol specifies the methods that will be used to undertake a specific systematic review. [11] defines the following components for a protocol:

1. Rationale, that is, the research reason.
2. Research questions the review intends to answer.
3. Strategy that will be used to search for primary studies, including search terms and sources for data extraction. This includes the formulation of keywords and the creation of search strings through their combination, as well as the selection of resources that will be searched, which may include indexed databases, specific journals and event annals.
4. Criteria for study inclusion and exclusion according to defined objectives such as the original language of the study and its area of application.
5. Checklists to evaluate the quality of the selected studies.
6. Strategy for data extraction. It will be used to define how the information required from each preliminary study will be obtained, i.e., whether it will be necessary to validate the data through some inference or manipulation.
7. Synthesis and analysis of the extracted data.
8. Project timeline to define the duration of each review step.

It should be noted that these components/steps are not sequential, and during the review process they can be executed more than once and undergo changes as a result of approval processes in both the planning and execution phases.

The protocol of systematic review represents the actual planning, since it describes all strategies, methods and considerations to be applied while it is executed. In order to support the protocol definition and the RS conduction, the software StArt (State of Art through Systematic Reviews), which has shown positive results, was used [26].

3.1 Research Questions

The objective of this systematic review is to identify the state of the art in algorithms used for automatic question classification, their advantages and disadvantages. There is a fair number of algorithms for text classification involving

[1] http://dl.acm.org

[2] http://www.scopus.com

[3] http://www.sciencedirect.com

[4] http://ieeexplore.ieee.org

[5] https://webofknowledge.com

[6] http://scholar.google.com

Machine Learning technology. In the case of questions, the amount of information is small, and their classification according to some labels become more complex. What is worse, there is also a lack of surveys on the state of the art of question classifiers, their possible applications and the quantity and quality of the available algorithms. Thus, the goal is to support the following primary research question:

What are the main automatic classification algorithms for questions?

Based on the primary question, specific questions of interest arise, as shown in Table 4:

TABLE 4
RESEARCH QUESTIONS

Research Question	Motivation
RQ1: What computational methods are used to implement classifiers?	This question identifies the methodology adopted in the construction of the main algorithms of Machine Learning to classify questions.
RQ2: Which taxonomies were adopted for classification?	The answer to this question makes it possible to identify the classification criteria applied.
RQ3: What are the main techniques used for feature extraction and selection?	The extraction and selection of relevant features have a direct impact on the performance of classifiers.
RQ4: What are the main instruments used to measure the classification results?	To check the different ways of analyzing the efficiency of the classification process.

3.2 Source and Study Selection

The selection criterion of archive sources considered the possibility of browsing studies on the web, regularly updated publications, availability of texts, quality of results, possibility of exporting bibliographic references, and search mechanism by title, abstract and keywords, all written in English. Formal and informal literature reviews written by experts and the opinion of researchers were also considered, both somehow involving the area of automatic text and question classification according to the guidelines described by [25].

Based on the requirements mentioned above, the following electronic databases were selected for the research: ACM Digital Library[1], Scopus[2], Elsevier Science Direct[3], IEEExplore Digital Library[4], Web of SciencS5 and Google Scholar[6].

Prior to the creation of the search string, a list of primary studies considered as control references was created. This list was extracted from the research work of the authors and also from secondary references on the subject previously analyzed, such as the work of [8], who developed an SR focused on question classification in computer exams, and the informal review published by [4], who considered the question classification in QA, IR, education and language conversion environments. The checklist is presented in Table 5:

TABLE 5
STUDY SELECTION FOR THE CONTROL GROUP

Ref.	Title
[22]	Exam Questions Classification Based on Bloom's Taxonomy Cognitive Level Using Classifiers Combination
[27]	A rule-based approach in Bloom's Taxonomy question classification through natural language processing
[28]	Bloom's taxonomy question categorization using rules and n-gram approach
[29]	An automatic classifier for exam questions in Engineering: A process for Bloom's taxonomy
[30]	Automated analysis of exam questions according to Bloom's taxonomy
[31]	Analyzing the cognitive level of classroom questions using machine learning techniques
[32]	Classification of high dimensional Educationa Data using Particle Swarm Classification

Based on the research questions and the SR guidelines, a search string was defined considering the areas and algorithms involved, that is, Machine Learning, Data Mining and Question Classification. Taking into account the possible synonyms, the following result was obtained:

Among the possible types of studies, the primary studies published in journals, conferences and book chapters, can be highlighted, preferably giving emphasis to the most recent publications in case similar studies are found. Based on the guidelines proposed by [11], Table 6 presents the inclusion/exclusion criteria defined for this SR:

```
((“MACHINE LEARNING”        OR
  “DATA MINING”             OR
  “TEXT MINING”             OR
  “DEEP LEARNING”)
                            AND

(“QUESTION CLASSIFICATION”  OR
 “QUESTION CLASSIFIER”      OR
 “QUESTION ANALYSIS”        OR
 “QUESTION ANALYZE”         OR
 “QUESTION CATEGORIZATION”))
```

TABLE 6
INCLUSION/EXCLUSION CRITERIA

Inclusion Criteria
IC 1: Primary studies
IC 2: Studies that present algorithms for question classification
IC 3: Studies published between January 2012 and July 2017
IC 4: Quality evaluation with score greater than or equal to 50
Exclusion Criteria
EC1: Secondary studies
EC 2: Incomplete studies or with few pages
EC 3: Studies written in a language other than English
EC 4: Studies with unavailable full text
EC 5: Studies focused on QA (*Question Answering*) or IR (*Information Retrieval*) areas.
EC 6: Studies focused on classification in other languages (e.g. Chinese, Persian, Hindi, Arabic)
EC 7: Studies that do not address algorithms for question classification

In the first stage of the process, called Preliminary Selection, the search string was adapted and executed for each electronic database of the source list. The initial search resulted in a selection of 4,460 compatible primary studies. Fig. 6 shows the steps for applying the inclusion/exclusion criteria, and the types of studies discarded are detailed on the right.

To limit the number of studies loaded, the database filter mechanism was used to select only the ones published after the beginning of January 2012 (IC3). This first filter resulted in 1,454 studies, which were identified and loaded into the StArt tool responsible for organizing and debugging information as well as deleting duplicate entries (176), resulting in 1,278 records.

The next step was the application of the inclusion criteria. At first, we read the title, keywords and area of knowledge addressed by the study to exclude the ones that did not meet the inclusion criteria and were not related to the research questions. After discarding 1.026 records, 252 primary studies remained for analysis. We read their abstract and again applied the inclusion/exclusion criteria, resulting in the exclusion of 92 more entries (details of the exclusion are presented in the second frame to the right in Fig. 6).

We fully retrieved the texts of 160 selected studies and read the introduction and conclusion sections. From this step, 53 other studies were discarded taking into consideration the inclusion/exclusion criteria, once the vast majority either had their questions translated into another language (24 studies, 45%) or not specifically addressed question classification (21 studies, 39%). Considering the bibliographical references of the selected studies and the secondary studies (reviews) discarded in the previous steps, 4 were added to the selection (see Fig. 7), resulting in 111 studies for the application of IC4 referring to the minimum score in the quality criteria. At last, 31 studies were eliminated, and the 80 remaining ones underwent a systematic review. The criteria with the lowest score were QC4 (description of limitations, mean = 0), QC7 (data availability, mean = 0.08), CQ5 (statistical analysis, mean = 0.3) and QC6 (validation test, mean = 0.4).

3.3 Quality Evaluation

The quality evaluation allows the selection of the most relevant studies with concrete results within the desired research theme. Five questions were obtained from the literature and two other proposals according to the scope and research questions were formulated.

The scoring scale was based on the dichotomy Yes (S)/ No (N), with a score of 1 for affirmative answers and 0 for negative ones, with the possibility of partial attendance of the question (P) and respective score of 0.5. As a minimum exclusion criterion, a score less than 3.5 was considered since it represented 50% of the utilization of the 7 possible scores (Table 7).

TABLE 7
QUESTIONS FOR QUALITY EVALUATION

Question	Allowed answers and scores
QC1: Is there a rationale to explain why the study was conducted? [33]	Y = 1, P = 0.5 and N = 0
QC 2: Is there a clear statement of the research objectives? [34], [35]	Y = 1, P = 0.5 and N = 0
QC 3: Is the proposed technique clearly described? [36]	Y = 1, P = 0.5 and N = 0
QC 4: Are the limitations of this study explicitly discussed? [37]	Y = 1, P = 0.5 and N = 0
QC 5: Were the results of the experiment reliably obtained through statistical analysis, for example? [35]	Y = 1, P = 0.5 and N = 0
QC 6: Does the study describe an experiment for algorithm validation?	Y = 1, P = 0.5 and N = 0
QC 7: Does the study provide the data for the experiment replication?	Y = 1, P = 0.5 and N = 0

Table 8 shows the selected studies based on the results of quality evaluation and the application of the inclusion criterion IC4. Analyzing the evaluation results, the average score of the studies was 4.43 (63.2%). The study S05 obtained the highest score (6.5 or 92.9%), while 13 studies achieved the threshold score (50%) for data extraction and result analysis (S4, S10, S13, S18, S21, S39, S41, S42, S45, S61, S63, S71 and S75). Taking into account the quality criteria individually, it can be seen that the criteria Q1 and Q2 obtained the highest scores, indicating well-founded rationales and clearly defined objectives; on the other hand, the criterion Q4 obtained a very low average (0.09), implying that most studies did not describe the possible limitations of their research.

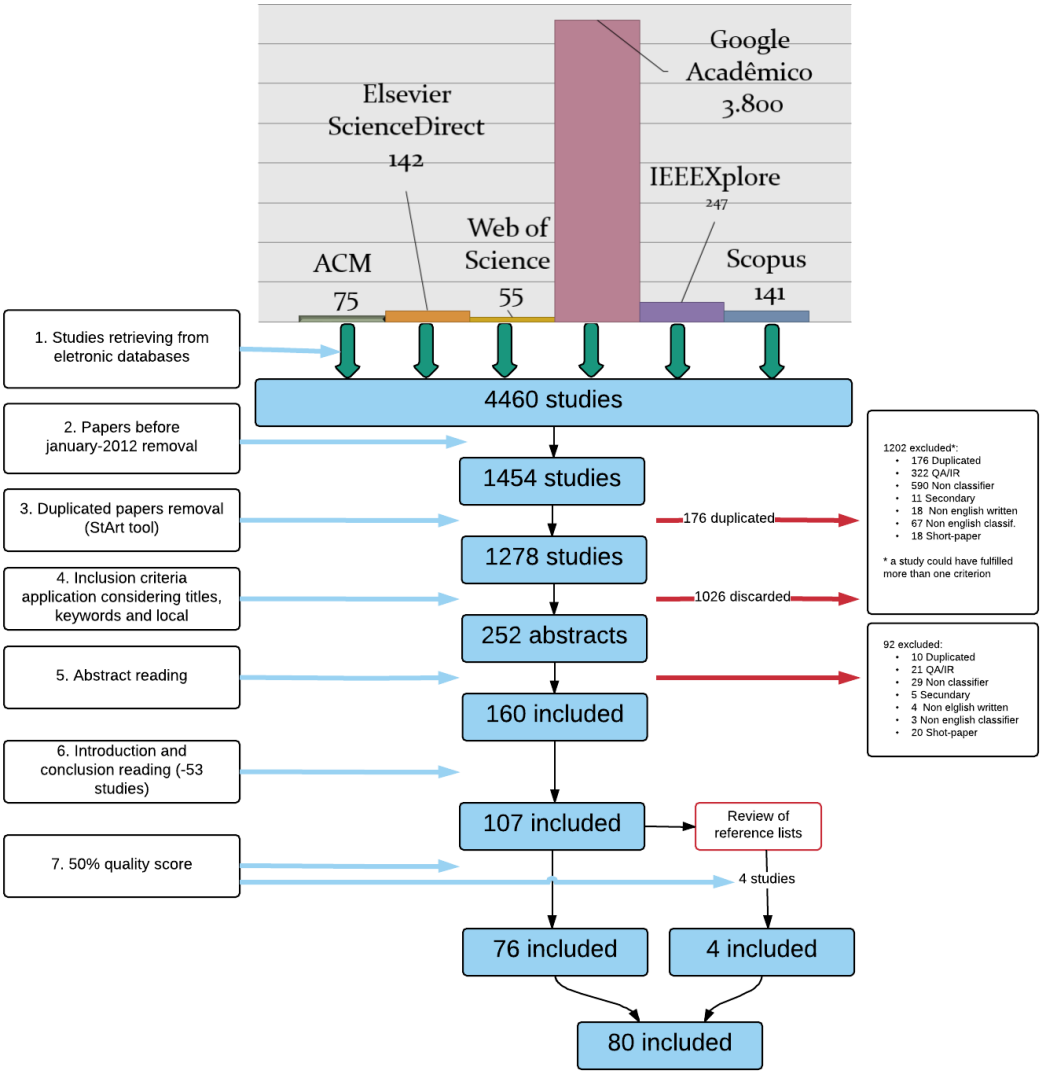

Fig. 6. Flow of application of the inclusion and exclusion criteria.

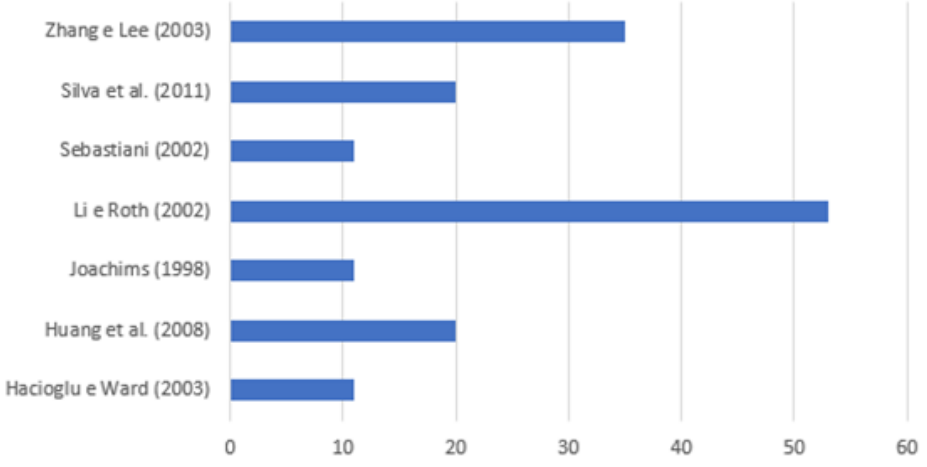

Fig. 7. Most cited references in the selected studies.

TABLE 8
STUDIES SELECTED THROUGH QUALITY EVALUATION

ID	Author	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Tt.	%
S01	Abduljabbar and Omar [22]	1.0	1.0	0.5	0.0	1.0	0.5	0.5	4.5	64.3
S02	Alahmadi [14]	1.0	1.0	1.0	0.0	1.0	1.0	0.5	5.5	78.6
S03	Bandyopadhyay et al. [38]	1.0	1.0	1.0	0.0	0.5	0.5	0.5	4.5	64.3
S04	Braum et al. [39]	1.0	1.0	0.5	0.0	0.5	0.5	0.0	3.5	50.0
S05	Chan et al. [40]	1.0	1.0	1.0	0.5	1.0	1.0	1.0	6.5	92.9
S06	Chen et al. [41]	1.0	1.0	1.0	0.0	0.5	0.5	0.5	4.5	64.3
S07	Chen et al. [42]	1.0	1.0	0.5	0.0	0.5	0.5	0.5	4.0	57.1
S08	Chernov et al. [43]	1.0	1.0	0.5	0.0	0.5	0.5	0.5	4.0	57.1
S09	Dalvi et al. [44]	1.0	1.0	0.5	0.0	0.5	0.5	0.5	4.0	57.1
S10	Diab et al. [45]	1.0	1.0	0.5	0.0	0.5	0.5	0.0	3.5	50.0
S11	Dubey et al. [46]	1.0	1.0	1.0	0.0	1.0	1.0	0.5	5.5	78.6
S12	Emmanuel et al. [19]	1.0	0.5	0.5	0.0	1.0	0.5	0.5	4.0	57.1
S13	Filice et al. [47]	1.0	0.5	0.5	0.0	0.5	0.5	0.5	3.5	50.0
S14	Foley et al. [48]	1.0	1.0	1.0	0.0	1.0	1.0	0.5	5.5	78.6
S15	Hao et al. [49]	1.0	1.0	1.0	0.0	0.5	0.5	0.5	4.5	64.3
S16	Hardy et al. [16]	1.0	1.0	1.0	0.5	1.0	1.0	0.5	6.0	85.7
S17	Haris et al. [28]	1.0	1.0	0.5	0.0	0.5	0.5	0.5	4.0	57.1
S18	Hoque et al. [50]	0.5	1.0	0.5	0.5	0.5	0.5	0.0	3.5	50.0
S19	Huang et al. [13]	1.0	1.0	1.0	0.0	0.5	0.5	0.5	4.5	64.3
S20	Huo et al. [51]	1.0	1.0	0.5	0.0	0.5	0.5	0.5	4.0	57.1
S21	Hutzler et al. [52]	1.0	0.5	1.0	0.0	0.5	0.5	0.0	3.5	50.0
S22	Jayakodi et al. [53]	1.0	1.0	1.0	0.5	0.5	0.5	0.5	5.0	71.4
S23	Jayakodi et al. [29]	1.0	1.0	1.0	0.0	0.5	0.5	0.0	4.0	57.1
S24	Jayakodi et al. [54]	0.5	1.0	0.5	0.5	0.5	0.5	0.5	4.0	57.1
S25	Kalchbrenner et al. [55]	1.0	1.0	1.0	0.0	0.5	0.5	0.5	4.5	64.3
S26	Karyawati et al. [56]	1.0	1.0	1.0	0.5	1.0	0.5	0.0	5.0	71,4
S27	Kim [57]	0.5	0.5	1.0	0.0	0.5	1.0	0.5	4.0	57.1
S28	Komninos et al. [58]	1.0	1.0	0.5	0.0	0.5	0.5	1.0	4.5	64.3
S29	La et al. [59]	1.0	1.0	1.0	0.0	1.0	1.0	0.5	5.5	78.6
S30	Le et al. [60]	1.0	1.0	1.0	0.0	0.5	0.5	0.5	4.5	64.3
S31	Lee et al. [61]	0.5	0.5	0.5	0.0	1.0	0.5	1.0	4.0	57.1
S32	Le-Hong et al. [62]	1.0	1.0	1.0	0.5	0.5	0.5	0.0	4.5	64.3
S33	Li and Roth [3]	1.0	1.0	0.5	0.0	0.5	0.5	1.0	4.5	64.3
S34	Li et al. [63]	1.0	1.0	0.5	0.0	0.5	0.5	0.5	4.0	57.1
S35	Lin et al. [64]	0.5	1.0	1.0	0.0	0.5	0.5	0.5	4.0	57.1
S36	Lu et al. [65]	1.0	1.0	1.0	0.0	0.5	0.5	0.0	4.0	57.1
S37	Luo et al. [66]	1.0	1.0	1.0	0.0	0.5	0.5	0.0	4.0	57.1
S38	Ma et al. [67]	0.5	1.0	1.0	0.0	0.5	0.5	1.0	4.5	64.3
S39	Madabushi et al. [68]	1.0	1.0	0.5	0.0	0.5	0.5	0.0	3.5	50.0
S40	Mahatme et al. [7]	1.0	1.0	0.5	0.0	0.5	0.5	0.5	4.0	57.1
S41	Marincic et al. [69]	0.5	0.5	0.5	0.0	0.5	0.5	1.0	3.5	50.0
S42	Mishra et al. [70]	1.0	1.0	0.5	0.0	0.5	0.5	0.0	3.5	50.0
S43	Mou et al. [71]	0.5	1.0	1.0	0.0	0.5	0.5	1.0	4.5	64.3
S44	Obasa et al. [72]	1.0	1.0	1.0	0.0	1.0	1.0	0.5	5.5	78.6
S45	Osadi et al. [73]	1.0	1.0	0.5	0.0	0.5	0.5	0.0	3.5	50.0
S46	Osman et al. [74]	1.0	1.0	1.0	0.0	0.5	0.5	0.0	4.0	57.1
S47	Osman et al. [75]	1.0	1.0	1.0	0.0	0.5	0.5	0.0	4.0	57.1
S48	Patrick e Li [76]	1.0	1.0	1.0	1.0	1.0	1.0	0.0	6.0	85.7
S49	Pillai et al. [77]	0.5	0.5	1.0	0.0	1.0	1.0	0.0	4.0	57.1
S50	Ping et al. [78]	1.0	1.0	1.0	0.0	1.0	1.0	0.5	5.5	78.6
S51	Pota et al. [17]	0.5	1.0	1.0	0.0	0.5	0.5	0.5	4.0	57.1
S52	Pota et al. [79]	1.0	1.0	1.0	0.0	0.5	0.5	0.5	4.5	64.3
S53	Poyraz et al. [80]	1.0	1.0	1.0	0.5	1.0	0.5	0.5	5.5	78.6
S54	Qi et al. [81]	1.0	1.0	0.5	0.0	0.5	0.5	0.5	4.0	57.1
S55	Qu et al. [82]	1.0	1.0	0.5	0.5	1.0	1.0	0.5	5.5	78.6
S56	Rahman et al. [83]	1.0	1.0	0.5	0.0	0.5	0.5	0.5	4.0	57.1
S57	Roberts et al. [84]	1.0	1.0	0.5	0.0	0.5	0.5	0.5	4.0	57.1
S58	Roberts et al. [85]	1.0	1.0	0.5	0.0	0.5	0.5	0.5	4.0	57.1
S59	Sangodiah et al. [86]	1.0	0.5	1.0	0.0	0.5	0.5	0.5	4.0	57.1
S60	Sarrouti et al. [87]	1.0	1.0	1.0	0.5	0.5	0.5	0.5	5.0	71.4
S61	Shanthi et al. [88]	1.0	1.0	0.5	0.0	0.5	0.5	0.0	3.5	50.0
S62	Silva et al. [89]	1.0	1.0	1.0	0.5	0.5	0.5	1.0	5.5	78.6
S63	Singh et al. [90]	1.0	1.0	0.5	0.0	0.5	0.5	0.0	3.5	50.0
S64	Tomas et al. [91]	1.0	1.0	1.0	0.5	1.0	1.0	0.5	6.0	85.7
S65	Van-Tu et al. [18]	1.0	1.0	1.0	0.0	0.5	0.5	0.5	4.5	64.3
S66	Wan et al. [92]	1.0	1.0	0.5	0.0	0.5	0.5	0.5	4.0	57.1
S67	Wang et al. [93]	1.0	1.0	1.0	0.5	1.0	1.0	0.5	6.0	85.7
S68	Wang et al. [94]	1.0	1.0	1.0	0.0	1.0	1.0	0.5	5.5	78.6
S69	Wang et al. [95]	1.0	1.0	1.0	0.0	1.0	1.0	0.5	5.5	78.6
S70	Wang et al. [96]	1.0	1.0	0.5	0.0	0.5	0.5	0.5	4.0	57.1
S71	Yadav et al. [97]	0.5	0.5	1.0	0.0	0.5	0.5	0.5	3.5	50.0
S72	Yahya et al. [98]	1.0	1.0	0.5	0.0	1.0	1.0	0.5	5.0	71.4
S73	Yahya et al. [32]	1.0	1.0	0.5	0.0	1.0	0.5	0.0	4.0	57.1
S74	Yoshikawa et al. [99]	1.0	1.0	0.5	0.0	1.0	1.0	1.0	5.5	78.6
S75	Zhang et al. [100]	1.0	1.0	0.5	0.0	0.5	0.5	0.0	3.5	50.0
S76	Zhang e Lee [101]	1.0	1.0	1.0	0.0	0.5	0.5	0.0	4.0	57.1
S77	Zhang et al. [102]	1.0	1.0	1.0	0.0	1.0	0.5	0.5	5.0	71.4
S78	Zhang et al. [103]	1.0	1.0	0.5	0.0	0.5	0.5	0.5	4.0	57.1
S79	Zhang et al. [104]	1.0	1.0	0.5	0.0	0.5	0.5	0.5	4.0	57.1
S80	Zhou et al. [105]	1.0	0.5	1.0	0.0	0.5	0.5	0.5	4.0	57.1
	Average	0.93	0.94	0.78	0.09	0.65	0.61	0.43	4.43	63.2

3.4 Data Extraction

Based on the guidelines proposed by [24] and the complete reading of 80 selected studies, a form was created to extract relevant information (Table 9). Basically, such information describes an overview of each study and details the answers to the proposed research questions.

The general information selected was: year, country, question or document database, study publication medium and sample size. Considering the research questions, the tabulated data were: algorithms (classifiers), classification criteria (taxonomies), preprocessing techniques, feature selection and indicators (metrics) used.

TABLE 9
FORM FOR DATA EXTRACTION

Information for Extraction	Description	Relevance
ID	Unique study identifier	General
Year	Year of publication	General
Country	Country of origin	General
Question Base	Question/document source	
Publication Medium	Journal, conference, workshop, book chapter and electronic archive	General
Sample Size	Number of questions used for the analysis of classifiers	General
Algorithms	Computational methods used, such as Artificial Neural Network. SVM. Naive Baves etc.	RQ1
Classification Criteria and Feature Selection	Taxonomies used for classification (e.g. Bloom)	RQ2
	Feature extraction with direct impact on the classification process (TF-IDF, Chi-Square Tests etc.)	RQ3
Result Metrics	Indicators to analyze accuracy, precision, recall and F1 Score	RQ4

4 Result Analysis

From the data extracted from the 80 selected studies, this chapter presents the results obtained considering its general information and proposed research questions.

4.1 Year of Publication

Despite establishing the years between 2012 and 2017 as the original observation interval of the studies, we decided to include in the bibliographic references the 4 most cited studies published between 2002 and 2011. Most of the studies were published in 2016 (28.75%), followed by the year 2015 (22.5%), 2014 (17.5%), 2012 (11.25%), 2013 (10%) and 2017 (5%). The punctual inclusions (years 2002, 2003, 2008 and 2011) correspond to 1.25% each (Fig. 8).

It is worth noting that 2017 is not finished yet and should have a considerable number of papers still to be published due to the increasing number of publications that has been observed since 2014.

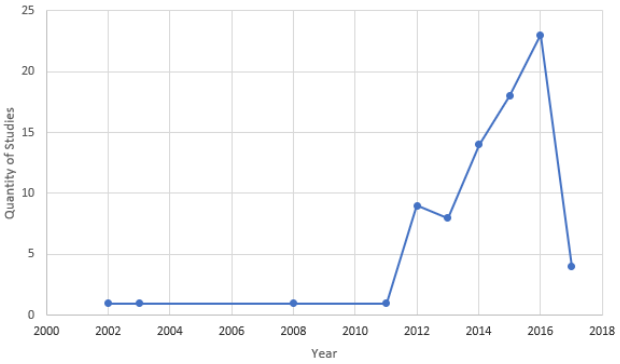

Fig. 8. Number of publications over the years.

4.2 Country

China (20%), United States (14%) and India (13%) concentrated nearly half (47%) of the primary studies selected. Taking into consideration the continents, Asia had 48 studies selected (60%), while Africa had less than 1%, with a single published work (Morocco). To complete the list, Europe represented 19% of the publications, followed by America (16%) and Oceania (4%). Fig. 9 shows the distribution of the studies among the countries with the highest number of publications:

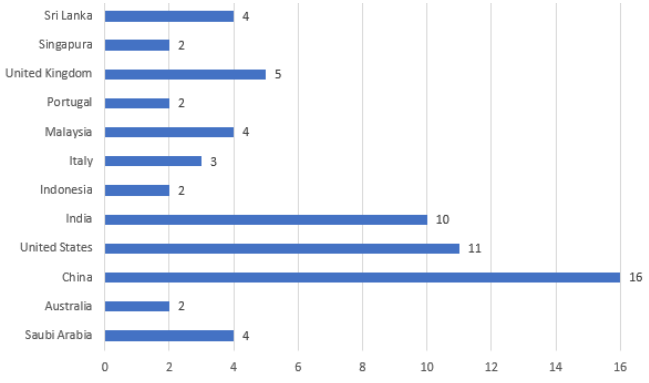

Fig. 9. Distribution of studies per country.

4.3 Question Base and Publication Medium

All 80 primary studies selected had their origin in the academic environment, which often resulted in the manual preparation of the question base using the evaluations produced inside the institution (20%) or the use of databases as reference in the question classification area, e.g. the database generated through the work of [3] (22.5%) carried out at the University of Illinois Urbana-Champaign (UIUC), in the United States, and provided by TREC (Text REtrieval Conference), one of the world's leading conferences on the subject. If we consider the non-exclusive use of the UIUC/TREC database, the use index rises to 33.75% of the studies.

In addition to the academic area, there are also several question sources available on the web, such as QA Yahoo! Answers (6.25%) and Stack Overflow (2.5%) systems, and bases with short text storage (similar to questions), such as

Reuters (6.25%), 20NewsGroups (6.25%) and Baidu (2.5%), which are also used together several times. A graph with the main question bases surveyed is shown in Fig. 10.

As already described in Table 9, the types of publication medium adopted in this work were journal, conference, workshop, book chapter and electronic archive. Book chapters include master's theses and Ph.D. dissertations. The electronic archive category was created to label studies stored in the arXiv.org electronic library of Cornell University not published elsewhere. Most of the studies were published in conferences (52.5%), journals (35%), electronic media (8.75%), book chapters (2.5%) and workshops (1.25%).

4.4 RQ1: Computational Methods

The purpose of this research question was to identify the different techniques used to implement classifiers. Two situations were observed: the first refers to the use of a simple algorithm or combination of algorithms to classify the questions; the other is about the main contribution of the study, in this case the focus on the classification logic (adaptation of basic algorithms, such as SVM and Neural Network, or creation of new algorithms), or on the feature identification, extraction and organization as a basis for the classifier mechanism.

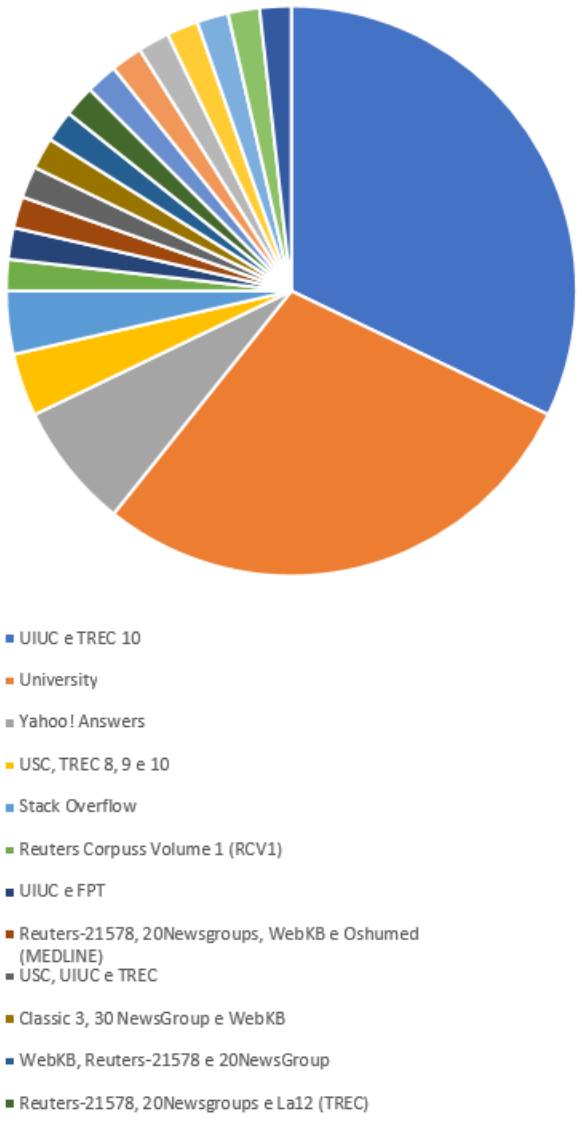

Fig. 10. Distribution of questions per question base.

Analyzing the computational methods of the 80 selected studies, 55 (68.75%) chose to use a single algorithm as a basis for the classifier. Table 10 shows the distribution of algorithm choice and the consequent predominance of SVM, Neural Network, Rules and Clustering.

Regarding the 25 studies that combined algorithms, we verified that SVM was also the preferred choice, followed by Naive Bayes, Rules, Decision Tree and Neural Network (Table 11).

TABLE 10
DISTRIBUTION OF UNITARY ALGORITHMS

Algorithm	Studies	Quantity	% (General)
SVM	S05, S08, S09, S12, S15, S36, S42, S48, S49, S50, S51, S52, S57, S59, S65	15	18.75%
Artificial Neural Network	S16, S25, S27, S30, S31, S38, S43, S69, S71, S77, S78, S80	12	15.00%
Rules	S10, S17, S22, S23, S24, S26, S39, S60	8	10.00%
Clustering	S11, S34, S40, S54, S75, S79	6	7.50%
kNN	S06, S66	2	2.50%
QACS	S35	1	1.25%
Linear Record	S61	1	1.25%
QPT	S03	1	1.25%
SMM	S74	1	1.25%
DC2	S64	1	1.25%
FSM	S18	1	1.25%
SDM	S14	1	1.25%
LMQC	S68	1	1.25%
SNoW	S33	1	1.25%
Naive Bayes	S53	1	1.25%
Decision Tree	S41	1	1.25%
Particle Swarm	S73	1	1.25%

TABLE 11
DISTRIBUTION OF ALGORITHMS IN STUDIES WITH COMBINATIONS

Algorithm	Studies	Qtty.	% (Partial)	% (General)
SVM	S01, S02, S04, S07, S19, S21, S28, S44, S45, S46, S47, S55, S56, S58, S62, S63, S67, S76	18	72.00%	41.25%
Naive Bayes	S01, S02, S20, S21, S32, S44, S45, S46, S47, S55, S76	11	44.00%	15.00%
Rules	S37, S44, S45, S58, S62, S63	6	24.00%	17.50%
Decision Tree	S02, S21, S44, S46, S47, S76	6	24.00%	8.75%
Artificial Neural Network	S20, S21, S28, S44, S56, S70	6	24.00%	22.50%
KNN	S01, S29, S45, S67, S76	5	20.00%	8.75%
Logistic Regression	S04, S37, S46, S47	4	16.00%	5.00%
Maximum Entropy	S19, S32, S55	3	12.00%	3.75%
SMO	S20, S44	2	8.00%	2.50%
SNoW	S76	1	4.00%	2.50%
Rochio	S72	1	4.00%	1.25%
RHC	S55	1	4.00%	1.25%
Clustering	S70	1	4.00%	8.75%
MPH	S55	1	4.00%	1.25%
S-EM	S07	1	4.00%	1.25%
Probability Estimation	S07	1	4.00%	1.25%
SPH	S55	1	4.00%	1.25%
PSO	S72	1	4.00%	1.25%
Adaboost	S29	1	4.00%	1.25%
Random Forest	S02	1	4.00%	1.25%

Finally, adding the 25 studies with algorithm combination to the initial analysis of 55 studies with unitary algorithms, SVM, Neural Network and Rules remained in the first three positions with 41.25%, 22.5% and 17.5%, respectively, while Naive Bayes occupied the forth place (15%).

As to the main contribution of the studies in the construction of question classifiers, 49 of them that focused on the classification logic were preferably chosen (61.25%) against 31 of those focused on the feature structuring (38.75%). As it can be seen in Fig. 11, algorithm combination was the main choice for selecting the studies, either regarding logic (12 studies, 15%) or feature (13 studies, 16.25%). Besides that, we could also verify that SVM had its main use in classifiers focused on feature extraction/selection (13 studies, 16.25%), while Neural Network (10 studies, 12.5%) and Rules (7 studies, 8.75%) were the main basic algorithms chosen for the new implementations of classification logic.

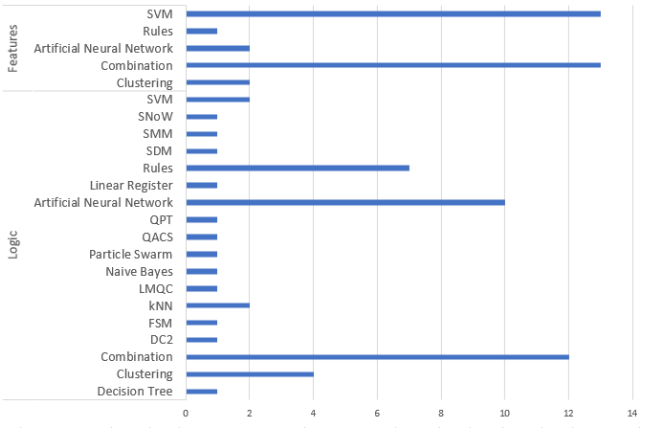

Fig. 11. Distribution of studies per construction of classifiers.

Results showed that Neural Network and SVM are the two types of classifiers mostly used to improve algorithms for classification. On the other hand, SVM was the main option when the choice was to make a combination of algorithms or to propose a new logic for feature handling.

The predominant use of Neural Network and SVM demonstrates that the processing cost has become an increasingly small barrier to the feasibility of research experiments as well as to the search for classifiers with the best possible performance indicators to classify questions.

4.5 RQ2: Taxonomies for Classification

The objective of this research question was to identify the most used criteria to classify questions and, therefore, the use of taxonomies already defined in the literature. It was observed that the use of authorial criteria of each research was predominant (35 studies, 43.75%, ranging from categories belonging to geographical areas and medical requirements to QA systems, such as Yahoo! Answers. It was also observed the use of binary classification (Yes/No, Positive/Negative, Relevant/Irrelevant) in 4 studies (5%). SVM was the most used algorithm (10 of 35 studies, 28.57%), followed by algorithm combination (9 of 35 studies, 25.71%).

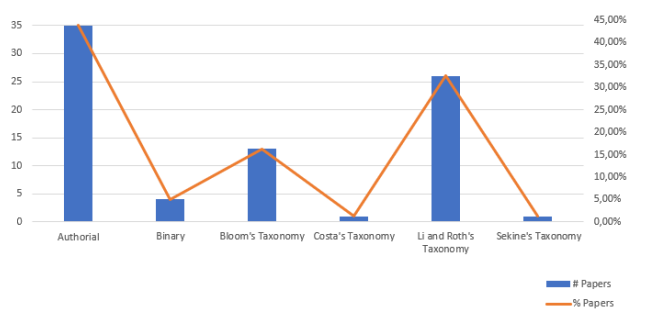

Fig. 12. Distribution of studies per criteria/taxonomy.

Among the taxonomies observed in Fig. 12, the one proposed by Li and Roth [3] was the most adopted (26 studies, 32.5%), followed by Bloom's taxonomy [106], which was used in 13 studies (16.25%), and taxonomies developed by Costa [88] and Sekine [91], each appearing in only 1 study (1.25%). Table 12 shows the identification of studies in each taxonomy/criterion found:

TABLE 12
IDENTIFICATION OF STUDIES IN EACH TAXONOMY

Taxonomy	Studies	Quantity	% (General)
Authorial	Other studies	35	43.75%
Binary	S44, S54, S56, S75	4	5.00%
Bloom	S01, S10, S11, S17, S22, S23, S24, S45, S46, S47, S59, S72, S73	13	16.25%
Costa	S61	1	1.25%
Li and Roth	S04, S08, S13, S16, S19, S25, S28, S30, S32, S33, S35, S38, S39, S43, S51, S52, S62, S63, S65, S69, S70, S71, S76, S77, S78, S80	26	32.50%
Sekine	S64	1	1.25%

Taking into account the relationship between the taxonomies developed either by Li and Roth or Bloom and the unitary algorithms implemented (no combinations nor authorial taxonomies), it was found that Rules were mostly used with Bloom's taxonomy (5 of 13 studies, 38.46%), while Neural Networks were preferably used with the taxonomy proposed by Lee and Roth (10 of 26 studies, 38.46%). This result reflects the compatibility between the use of mapping Rules and the learning levels proposed by Bloom; it also shows the ease of adapting Neural Networks to hierarchical classifications as proposed by Lee and Roth.

4.6 RQ3: Feature Extraction and Selection

The objective of this research question was to identify which mechanisms were used for feature extraction after the initial preprocessing of questions, as well as which techniques were responsible for assigning weights or indexes to features in order to select the most relevant ones to a certain question. After analyzing the 80 studies selected, it was possible to note that the majority chose either the extraction or the selection technique, which means that only 36% of the studies analyzed (29 studies) described both the extraction and the selection of features used to classify questions. Individually, extraction appeared in 62 studies (77.5%) and selection in 45 studies (56.25%), while only 9 studies (11,25%) did not present the extraction nor the selection technique.

TABLE 13
STUDIES PER TYPE OF FEATURE EXTRACTION

Type of Extraction	Studies	Quantity	% (General)
BOW	S02, S04, S08, S13, S26, S32, S41, S42, S44, S46, S47, S48, S51, S52, S55, S57, S58, S59, S62, S63, S64, S65, S69, S74	24	30.00%
N-grams	S04, S05, S08, S17, S19, S25, S34, S37, S38, S42, S46, S47, S55, S65, S70, S76, S80	17	21.25%
Head words	S16, S18, S19, S39, S51, S52, S59, S62, S63, S65, S71	11	13.75%
POS-tags	S05, S08, S17, S33, S45, S46, S47, S58, S59, S69, S71	11	13.75%
Wh-words	S16, S19, S32, S39, S42, S44, S51, S52, S63	9	11.25%
Bigrams	S41, S48, S51, S58, S63, S64	6	7.50%
Entities	S26, S33, S60, S68, S69, S71	6	7.50%
Question words	S41, S42, S48, S61, S71	5	6.25%
Hypernyms/Hiponyms	S15, S16, S19, S62, S65	5	6.25%
Verbs	S39, S51, S52, S63	4	5.00%
Nouns	S51, S52, S63	3	3.75%
Synsets	S56, S62	2	2.50%
Keywords	S05, S18	2	2.50%
Multi-Head words	S51, S52	2	2.50%
Trigrams	S63, S64	2	2.50%
Related words	S63, S65	2	2.50%
Skipgrams	S28, S69	2	2.50%
Adjectives	S39, S52	2	2.50%
Word Vector	S31, S45	2	2.50%
Concepts	S68	1	1.25%
Taxonomy-based	S59	1	1.25%
CBOW	S27	1	1.25%
Chunks	S33	1	1.25%
Tree Kernel	S76	1	1.25%
Word Segmentation	S54	1	1.25%
What-WDT	S58	1	1.25%
After-How	S52	1	1.25%
KCRF	S29	1	1.25%
Question Mark	S44	1	1.25%
TOPSIS	S41	1	1.25%
Question Patterns	S65	1	1.25%
Tree Tags	S58	1	1.25%
Focus-words	S08	1	1.25%
Syntactic Maps	S39	1	1.25%
Focus-words lemmas	S08	1	1.25%
What-WP	S58	1	1.25%
Syntactic Trees	S05	1	1.25%
Bag-of-concepts	S02	1	1.25%
Clue Words	S34	1	1.25%
Question Categories	S65	1	1.25%
Question Expansion	S65	1	1.25%

TABLE 14
STUDIES PER TYPE OF FEATURE SELECTION

Type of Selection	Studies	Quantity	% (General)
TF-IDF	S04, S05, S17, S44, S46, S47, S49, S54, S56, S79	10	12.50%
Cosine Similarity	S05, S22, S23, S24, S28, S68	6	7.50%
TF	S44, S53, S72, S73	4	5.00%
Binary Vector	S20, S44, S59	3	3.75%
Chi-Square	S01, S36, S44	3	3.75%
Word2vec	S27, S69	2	2.50%
LDA	S11, S66	2	2.50%
Action Verbs Similarity	S10	1	1.25%
SNOMED	S48	1	1.25%
QF-ICF	S49	1	1.25%
Feature Graph	S25	1	1.25%
CCE	S50	1	1.25%
Gain Ratio	S44	1	1.25%
PIF - Positive Impact Factor	S12	1	1.25%
GloVe	S69	1	1.25%
Semantically Related Words	S33	1	1.25%
ICF	S67	1	1.25%
Symmetrical Uncertainty	S44	1	1.25%
Information Gain	S44	1	1.25%
Yahoo! PlaceMaker	S07	1	1.25%
IQF-QF-ICF	S49	1	1.25%
PAS	S48	1	1.25%
KCRF	S29	1	1.25%
Probability	S68	1	1.25%
KFM	S18	1	1.25%
Relatedeness	S68	1	1.25%
Term-Weighting Scheme	S09	1	1.25%
SMO wrapper	S44	1	1.25%
BM25	S75	1	1.25%
SVDRD	S50	1	1.25%
Typed Dependencies	S32	1	1.25%
DC2	S64	1	1.25%
Word Embedding	S28	1	1.25%
Apriori	S06	1	1.25%
TF-ICF	S67	1	1.25%
Levenshtein Distance	S58	1	1.25%
VRF	S49	1	1.25%
MDS	S35	1	1.25%
Word Ranking	S57	1	1.25%
Mutual Information	S01	1	1.25%
WSD	S16	1	1.25%
N2WET	S50	1	1.25%
NWNET	S50	1	1.25%
Odds Ratio	S01	1	1.25%

TABLE 15
METRICS USED IN THE STUDIES

Metric	Studies	Qtty	% (General)
Accuracy	S02, S06, S09, S10, S12, S13, S15, S16, S18, S19, S20, S22, S23, S24, S25, S27, S28, S30, S31, S32, S33, S34, S35, S36, S38, S39, S41, S42, S43, S45, S48, S49, S50, S51, S52, S53, S54, S57, S59, S60, S61, S62, S63, S64, S65, S68, S69, S70, S71, S74, S75, S76, S77, S78, S80	55	68.75%
Precision	S01, S02, S03, S04, S06, S07, S08, S10, S11, S16, S17, S18, S19, S26, S29, S44, S45, S46, S47, S48, S49, S56, S57, S58, S64, S66, S68, S72, S79	29	36.25%
Recall	S01, S02, S03, S04, S06, S07, S10, S11, S16, S17, S18, S19, S26, S29, S44, S45, S46, S47, S48, S49, S56, S57, S58, S66, S68, S72, S79	27	33.75%
F1	S01, S03, S04, S06, S07, S10, S13, S17, S29, S44, S45, S46, S47, S48, S49, S53, S55, S57, S58, S66, S68, S72, S75, S79	24	30.00%
Macro F1	S02, S05, S07, S37, S50, S55, S58, S67, S73	9	11.25%
Micro F1	S02, S05, S07, S37, S50, S55, S58, S67	8	10.00%
Processing Time	S09, S12, S35, S61	4	5.00%
Standard Deviation	S16, S21, S53	3	3.75%
Error Rate	S10, S21, S34	3	3.75%
T Test	S64, S69	2	2.50%
Total Hits	S46, S47	2	2.50%
Correlation	S46, S47	2	2.50%
Kappa	S46, S47	2	2.50%
M-Average	S10, S26	2	2.50%
Target Hits	S46, S47	2	2.50%
Macro Accuracy	S64	1	1.25%
P@1	S14	1	1.25%
NDCg	S14	1	1.25%
Macro Recall	S37	1	1.25%
Gain	S50	1	1.25%
RR	S14	1	1.25%
Micro Accuracy	S64	1	1.25%
Overgeneration	S26	1	1.25%
Efficiency	S61	1	1.25%
P@10	S26	1	1.25%
Perplexity	S36	1	1.25%
Kernel gamma	S49	1	1.25%
Average Sysset	S23	1	1.25%
Mean Distance	S21	1	1.25%
Macro Precision	S37	1	1.25%
Boolean Success Rate	S21	1	1.25%
Significance test	S76	1	1.25%
Micro Precision	S37	1	1.25%
Undergeneration	S26	1	1.25%
Micro Recall	S37	1	1.25%
MRR	S26	1	1.25%
Not informed	S40	1	1.25%

As to different mechanisms for extracting features (Table 13), the BOW (Bag-of-words) technique was the most used (24 studies, 30%), followed by N-Grams (17 studies, 21.25%) and Headwords and POS-tags, with 11 studies each (13.75%). On the other hand, the feature selection appeared in 45 studies (56.25%). In general, we observed the great predominance of N-Grams and their variations (BOW, Bigrams, Trigrams), representing 61.25% of the studies, which means that the use of sequential tokens extracted from the question prevailed over the other forms involving concepts/meanings, keywords, or relationships between the question terms.

In the case of feature selection (Table 14), the TF-IDF technique was the most used (10 studies, 12.5%), followed by Cosine Similarity (6 studies, 7.5%), TF (4 studies, 5%), Binary Vector and Chi-Square, both with 3 studies each (3.75%). This result demonstrates a "spraying" in the use of selection techniques with a slight TF-IDF domain.

4.7 QP4: Result Metrics

Accuracy, Precision, Recall and F1 (F-measure) are result indicators widely used in academic studies and were also corroborated in this SR (see all results in Table 15). Accuracy was the most employed metric, appearing in 55 of 80 studies, followed by Precision (29 studies, 36.25%), Recall (27 studies, 33.65%) and F1 (24 studies, 12.12%).

5 Final Consideration

The focus of this Systematic Review was to achieve the state-of-the-art in the techniques, criteria and indicators associated with question classification. This article surveyed studies directly and specifically involved in question classification and disregarded works with broader scope, such as text classification (which encompasses the questions) or associated scope but different focus (QA, IR and language translation systems), e.g. document handling, system architecture, or expressions in a particular language.

Eighty studies were selected according to inclusion, exclusion and quality criteria, in which two main computational methods, 6 taxonomies/criteria, 2 mechanisms of feature control and the 4 most used indicators to measure question classification were identified.

Between the two types of computational methods used, the great majority of studies opted for contributions in the classification logic. Neural Networks, SVM and algorithm combination represented the most used computational methods.

Feature handling stood out both in algorithm identification and selection and extraction techniques, since more than a third of the studies (38.75%) implemented new algorithms for such handling, and almost 90% (88.75%) used some extraction/selection mechanism.

The criteria/taxonomies and the indicators did not show any newness in relation to what had already been described in the literature (sections 2.5 and 2.6): the use of categories defined by the authors had the same prominence as did Accuracy/Precision/Recall/F1-score, which proved to be the most adopted metrics to measure the performance of the classifiers.

As threats to validity and limitations, we can cite the use of few researchers in the survey and the possible adoption of subjective decisions regarding the inclusion/exclusion/quality criteria applied to the selected studies. These factors may have directly influenced the absence of some search source, resulting in the non-inclusion of some important studies and/or exclusion of studies relevant to the research objectives.

In relation to future work, our purpose is to carry out a detailed survey based on the quality analysis produced in this SR to retrieve data whose sources were made available in their respective studies. With this, the intention is to conduct a meta-analysis of such studies to compare and validate the disclosed results.

References

[1] S. M. Weiss, N. Indurkhya e T. Zhang, Fundamentals of predictive text mining, vol. 41, Springer, 2010.

[2] C. C. Aggarwal e C. Zhai, Mining text data, Springer Science \& Business Media, 2012.

[3] X. Li e D. Roth, "Learning question classifiers," em *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, 2002.

[4] A. Sangodiah, A. Muniandy e L. E. Heng, "Question classification using statistical approach: a complete review," *Journal of Theoretical and Applied Information Technology,* vol. 71, nº 3, pp. 386-395, 2015.

[5] C. Romero e S. Ventura, "Data mining in education," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* vol. 3, nº 1, pp. 12-27, 2013.

[6] N. Sonwalkar, "Adaptive individualization: The next generation of online education," *On the horizon,* vol. 16, nº 1, pp. 44-47, 2008.

[7] V. P. Mahatme e K. Bhoyar, "Questions Categorization in E-Learning Environment Using Data Mining Technique," *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering,* vol. 10, nº 1, pp. 93-97, 2016.

[8] M. K. TAQI e R. ALI, "AUTOMATIC QUESTION CLASSIFICATION MODELS FOR COMPUTER PROGRAMMING EXAMINATION: A SYSTEMATIC LITERATURE REVIEW," *Journal of Theoretical and Applied Information Technology,* vol. 93, nº 2, 2016.

[9] S. Jayalakshmi e A. Sheshasaayee, "Question classification: A review of state-of-the-art algorithms and approaches," *Indian Journal of Science and Technology,* vol. 8, nº 29, 2015.

[10] A. Patra e D. Singh, "A survey report on text classification with different term weighing methods and comparison between classification algorithms," *International Journal of Computer Applications,* vol. 75, nº 7, 2013.

[11] B. Kitchenham, "Procedures for performing systematic reviews," *Keele, UK, Keele University,* vol. 33, nº 2004, pp. 1-26, 2004.

[12] J. Biolchini, P. G. Mian, A. C. C. Natali e G. H. Travassos, "Systematic review in software engineering," *System Engineering and Computer Science Department COPPE/UFRJ, Technical Report ES,* vol. 679, nº 05, p. 45, 2005.

[13] Z. Huang, M. Thint e Z. Qin, "Question classification using head words and their hypernyms," em *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008.

[14] A. Alahmadi, "Automatic text classification using bag of words and bag of concepts based representations," 2016.

[15] S. Jayalakshmi e A. Sheshasaayee, "INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES \& RESEARCH TECHNOLOGY A REVIEW ON QUESTION CLASSIFICATION USING MACHINE LEARNING BASED ON SEMANTIC FEATURES," 2015.

[16] H. a. C. Y.-N. Hardy, "Question classification using extreme learning machine on semantic features," *Journal of ICT Research and Applications,* vol. 7, nº 1, pp. 36-58, 2013.

[17] M. Pota, A. Fuggi, M. Esposito e G. De Pietro, "Extracting compact sets of features for question classification in cognitive systems: a comparative study," em *P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2015 10th International Conference on*, 2015.

[18] N. Van-Tu e L. Anh-Cuong, "Improving question classification by feature extraction and selection," *Indian Journal of Science and Technology,* vol. 9, nº 17, 2016.

[19] M. Emmanuel, S. M. Khatri e D. R. Babu, "A Novel scheme for Term Weighting in Text Categorization: Positive Impact factor," em *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*, 2013.

[20] F. P. Shah e V. Patel, "A review on feature selection and feature extraction for text classification," em *Wireless Communications, Signal Processing and Networking (WiSPNET), International Conference on*, 2016.

[21] F. S. Gharehchopogh e Y. Lotfi, "Machine learning based question classification methods in the question answering systems," *Int J Innovat Appl Stud,* vol. 4, nº 2, 2013.

[22] D. Abduljabbar e N. Omar, "Exam questions classification based on Bloom's taxonomy cognitive level using classifiers combination," *Journal of Theoretical and Applied Information Technology,* vol. 78, nº 3, pp. 447-455, 2015.

[23] Wikipedia, *F1 Score,* 2017.

[24] S. Keele e others, "Guidelines for performing systematic literature reviews in software engineering," em *Technical report, Ver. 2.3 EBSE Technical Report. EBSE*, sn, 2007.

[25] E. Y. Nakagawa, K. R. F. Scannavino, S. C. P. F. Fabbri e F. C. Ferrari, Revis{\~a}o Sistem{\'a}tica da Literatura em Engenharia de Software: Teoria e Pr{\'a}tica, Elsevier Brasil, 2017.

[26] E. Hernandes, A. Zamboni, S. Fabbri e A. D. Thommazo, "Using GQM and TAM to evaluate StArt-a tool that supports Systematic Review," *CLEI Electronic Journal,* vol. 15, nº 1, pp. 3-3, 2012.

[27] S. S. Haris e N. Omar, "A rule-based approach in Bloom's Taxonomy question classification through natural language processing," em *Computing and Convergence Technology (ICCCT), 2012 7th International Conference on*, 2012.

[28] S. S. Haris e N. Omar, "BLOOM'S TAXONOMY QUESTION CATEGORIZATION USING RULES AND N-GRAM APPROACH.," *Journal of Theoretical \& Applied Information Technology,* vol. 76, nº 3, 2015.

[29] K. Jayakodi, M. Bandara e I. Perera, "An automatic classifier for exam questions in Engineering: A process for Bloom's taxonomy," em *Teaching, Assessment, and Learning for Engineering (TALE), 2015 IEEE International Conference on*, 2015.

[30] N. Omar, S. S. Haris, R. Hassan, H. Arshad, M. Rahmat, N. F. A. Zainal e R. Zulkifli, "Automated analysis of exam questions according to Bloom's taxonomy," *Procedia-Social and Behavioral Sciences,* vol. 59, pp. 297-303, 2012.

[31] A. A. Yahya, A. Osman, A. Taleb e A. A. Alattab, "Analyzing the cognitive level of classroom questions using machine learning techniques," *Procedia-Social and Behavioral Sciences,* vol. 97, pp. 587-595, 2013.

[32] A. A. Yahya e A. Osman, "Classification of high dimensional Educational Data using Particle Swarm Classification," em *Computer Systems and Applications (AICCSA), 2014 IEEE/ACS 11th International Conference on*, 2014.

[33] S. Mahdavi-Hezavehi, M. Galster e P. Avgeriou, "Variability in quality attributes of service-based software systems: A systematic literature review," *Information and Software Technology,* vol. 55, nº 2, pp. 320-343, 2013.

[34] T. Dyba e T. Dingsoyr, "Empirical studies of agile software development: A systematic review," *Information and software technology,* vol. 50, nº 9, pp. 833-859, 2008.

[35] N. Salleh, E. Mendes e J. Grundy, "Empirical studies of pair programming for CS/SE teaching in higher education: A systematic literature review," *IEEE Transactions on Software Engineering,* vol. 37, nº 4, pp. 509-525, 2011.

[36] P. a. S. A. a. I. R. a. M. M. N. Achimugu, "A systematic literature review of software requirements prioritization research," *Information and software technology,* pp. 568-585, 2014.

[37] W. a. L. P. a. T. A. a. V. V. H. Ding, "Knowledge-based approaches in software documentation: A systematic literature review," *Information and Software Technology,* pp. 545-567, 2014.

[38] S. B. S. Bandyopadhyay, "Question Classification and Answering from Procedural Text in English," em *24th International Conference on Computational Linguistics*, 2012.

[39] H. M. Braum, S. J. Rigo e J. L. Barbosa, "MODELO DE CLASSIFICA{\c{C}}{\~A}O AUTOM{\'A}TICA DE QUEST{\~O}ES NA L{\'I}NGUA PORTUGUESA," *RENOTE,* vol. 12, nº 2.

[40] W. Chan, W. Yang, J. Tang, J. Du, X. Zhou e W. Wang, "Community question topic categorization via hierarchical kernelized classification," em *Proceedings of the 22nd ACM international conference on Conference on information \& knowledge management*, 2013.

[41] C. Chen, H. Han e Z. Liu, "KNN question classification method based on Apriori algorithm," 2014.

[42] L. Chen, D. Zhang e M. Levene, "Identifying local questions in community question answering," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* vol. 7675 LNCS, pp. 306-315, 2012.

[43] A. Chernov, V. Petukhova e D. Klakow, "Linguistically motivated question classification," em *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania*, 2015.

[44] M. Dalavi e S. Cheke, "Hadoop MapReduce implementation of a novel scheme for term weighting in text categorization," em *Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014 International Conference on*, 2014.

[45] S. Diab e B. Sartawi, "Classification of Questions and Learning Outcome Statements (LOS) Into Blooms Taxonomy (BT) By Similarity Measurements Towards Extracting Of Learning Outcome from Learning Material," *arXiv preprint arXiv:1706.03191,* 2017.

[46] M. Dubey e V. Goyal, "Classifying stack overflow questions based on bloom's taxonomy," 2016.

[47] S. Filice, D. Croce e R. Basili, "A Stratified Strategy for Efficient Kernel-Based Learning.," em *AAAI*, 2015.

[48] J. Foley e J. Allan, "Retrieving hierarchical syllabus items for exam question analysis," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* vol. 9626, pp. 575-586, 2016.

[49] T. Hao, W. Xie e F. Xu, "A WordNet Expansion-Based Approach for Question Targets Identification and Classification," em *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, Springer, 2015, pp. 333-344.

[50] M. M. Hoque, T. Goncalves e P. Quaresma, "Classifying Questions in Question Answering System Using Finite State Machines with a Simple Learning Approach," *Sponsors: National Science Council, Executive Yuan, ROC Institute of Linguistics, Academia Sinica NCCU Office of Research and Development,* p. 409, 2013.

[51] L.-F. Huo, L.-M. Zhang e X.-Q. Zhao, "Question recognition based on subject," *Lecture Notes in Electrical Engineering,* vol. 375, pp. 353-361, 2016.

[52] D. Hutzler, E. David, M. Avigal e R. Azoulay, "Learning Methods for Rating the Difficulty of Reading Comprehension Questions," em *Software Science, Technology and Engineering (SWSTE), 2014 IEEE International Conference on*, 2014.

[53] K. Jayakodi, M. Bandara e D. Meedeniya, "An automatic classifier for exam questions with WordNet and Cosine similarity," em *Moratuwa Engineering Research Conference (MERCon), 2016*, 2016.

[54] K. Jayakodi, M. Bandara, I. Perera e D. Meedeniya, "WordNet and Cosine Similarity based Classifier of Exam Questions using Bloom's Taxonomy.," *International Journal of Emerging Technologies in Learning,* vol. 11, nº 4, 2016.

[55] N. Kalchbrenner, E. Grefenstette e P. Blunsom, "A convolutional neural network for modelling sentences," *arXiv preprint arXiv:1404.2188,* 2014.

[56] A. E. Karyawati, E. Winarko, A. Azhari e A. Harjoko, "Ontology-based why-question analysis using lexico-syntactic patterns," *International Journal of Electrical and Computer Engineering,* vol. 5, nº 2, p. 318, 2015.

[57] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882,* 2014.

[58] A. Komninos e S. Manandhar, "Dependency based embeddings for sentence classification tasks," em *Proceedings of NAACL-HLT*, 2016.

[59] L. La, Q. Guo, D. Yang e Q. Cao, "Multiclass boosting with adaptive group-based kNN and its application in text categorization," *Mathematical Problems in Engineering,* vol. 2012, 2012.

[60] P. Le e W. Zuidema, "The Forest Convolutional Network: Compositional Distributional Semantics with a Neural Chart and without Binarization.," em *EMNLP*, 2015.

[61] J. Y. Lee e F. Dernoncourt, "Sequential short-text classification with recurrent and convolutional neural networks," *arXiv preprint arXiv:1603.03827,* 2016.

[62] P. Le-Hong, X.-H. Phan e T.-D. Nguyen, "Using dependency analysis to improve question classification," em *Knowledge and Systems Engineering*, Springer, 2015, pp. 653-665.

[63] Y. Li, A. Tripathi e A. Srinivasan, "Challenges in Short Text Classification: The Case of Online Auction

Disclosure," 2016.

[64] Z. Lin, H. Wang e S. McClean, "Tree Similarity Measurement for Classifying Questions by Syntactic Structures," em *International Conference on Intelligent Computing*, 2016.

[65] Z. Lu, Y.-R. Lin, Q. Zhang e M. Chen, "Classifying Questions into Fine-Grained Categories Using Topic Enriching," em *Information Reuse and Integration (IRI), 2016 IEEE 17th International Conference on*, 2016.

[66] Y. Luo, T. F. Boucher, T. Oral, D. Osofsky e S. Weber, "A Study on Expert Sourcing Enterprise Question Collection and Classification.," em *LREC*, 2014.

[67] M. Ma, L. Huang, B. Xiang e B. Zhou, "Dependency-based convolutional neural networks for sentence embedding," *arXiv preprint arXiv:1507.01839,* 2015.

[68] H. T. Madabushi e M. Lee, "High Accuracy Rule-based Question Classification using Question Syntax and Semantics," 2016.

[69] D. Marincic, T. Kompara e M. Gams, "Question classification with active learning," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* vol. 7499 LNAI, pp. 673-680, 2012.

[70] S. K. Mishra, P. Kumar e S. K. Saha, "A Support Vector Machine Based System for Technical Question Classification," em *International Conference on Mining Intelligence and Knowledge Exploration*, 2015.

[71] L. Mou, H. Peng, G. Li, Y. Xu, L. Zhang e Z. Jin, "Tree-based convolution: A new neural architecture for sentence modeling," em *Proceedings of Conference on Empirical Methods in Natural Language Processing (to appear)*, 2015.

[72] A. I. Obasa, N. Salim e A. Khan, "Hybridization of Bag-of-Words and Forum Metadata for Web Forum Question Post Detection," *Indian Journal of Science and Technology,* vol. 8, nº 32, 2016.

[73] K. Osadi, M. Fernando e W. Welgama, "Ensemble Classifier based Approach for Classification of Examination Questions into Bloom's Taxonomy Cognitive Levels," 2017.

[74] A. OSMAN e A. A. YAHYA, "CLASSIFICATIONS OF EXAM QUESTIONS USING NATURAL LANGUAGE SYNTATIC FEATURES: A CASE STUDY BASED ON BLOOM'S TAXONOMY".

[75] A. OSMAN e A. A. YAHYA, "CLASSIFICATIONS of EXAM QUESTIONS USING LINGUISTICALLY-MOTIVATED FEATURES: A CASE STUDY BASED on BLOOM'S TAXONOMY".

[76] J. Patrick e M. Li, "An ontology for clinical questions about the contents of patient notes," *Journal of Biomedical Informatics,* vol. 45, nº 2, pp. 292-306, 2012.

[77] P. G. Pillai e J. Narayanan, "Question categorization using SVM based on different term weighting methods," *International Journal on Computer Science and Engineering,* vol. 4, nº 5, p. 938, 2012.

[78] Y. PING, Y. jian ZHOU, C. XUE e Y. xian YANG, "Efficient representation of text with multiple perspectives," *The Journal of China Universities of Posts and Telecommunications ,* vol. 19, nº 1, pp. 101-111, 2012.

[79] M. Pota, M. Esposito e G. De Pietro, "A forward-selection algorithm for SVM-based question classification in cognitive systems," *Smart Innovation, Systems and Technologies,* vol. 55, pp. 587-598, 2016.

[80] M. Poyraz, Z. H. Kilimci e M. C. Ganiz, "Higher-order smoothing: a novel semantic smoothing method for text classification," *Journal of Computer Science and Technology,* vol. 29, nº 3, pp. 376-391, 2014.

[81] X. Qi, L. Su, B. Yang, J. Chen, Y. Li e J. Liu, "Question Classification Based on Hadoop Platform," em *International Conference on Cloud Computing*, 2014.

[82] B. Qu, G. Cong, C. Li, A. Sun e H. Chen, "An evaluation of classification models for question topic categorization," *Journal of the American Society for Information Science and Technology,* vol. 63, nº 5, pp. 889-903, 2012.

[83] A. K. M. M. M. Rahman e C. K. Roy, "Embedded Emotion-based Classification of Stack Overflow Questions Towards the Question Quality Prediction," 2016.

[84] K. Roberts, H. Kilicoglu, M. Fiszman e D. Demner-Fushman, "Automatically classifying question types for consumer health questions," *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium,* vol. 2014, pp. 1018-1027, 2014.

[85] K. Roberts, H. Kilicoglu, M. Fiszman e D. Demner-Fushman, "Decomposing consumer health questions," em *BioNLP Workshop*, 2014.

[86] A. SANGODIAH, R. AHMAD, W. AHMAD e W. FATIMAH, "TAXONOMY BASED FEATURES IN QUESTION CLASSIFICATION USING SUPPORT VECTOR MACHINE.," *Journal of Theoretical \& Applied Information Technology,* vol. 95, nº 12, 2017.

[87] M. Sarrouti, A. Lachkar e S. E. A. Ouatik, "Biomedical question types classification using syntactic and rule based approach," em *Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference on*, 2015.

[88] P. Shanthi e I. Krishnamurthi, "A Semantic Approach for Question Classification Using Register Linear Based Model," *Middle-East Journal of Scientific Research,* vol. 23, nº 4, pp. 685-694, 2015.

[89] J. Silva, L. Coheur, A. C. Mendes e A. Wichert, "From symbolic to sub-symbolic information in question classification," *Artificial Intelligence Review,* vol. 35, nº 2, pp. 137-154, 2011.

[90] V. Singh e S. K. Dwivedi, "{An Integrated Pattern Matching and Machine Learning Approach for Question Classification}," em *{2015 1st International Conference on Next Generation Computing Technologies (NGCT)}*, {2015}.

[91] D. Tomas e J. L. Vicedo, "{Minimally supervised question classification on fine-grained taxonomies}," *{KNOWLEDGE AND INFORMATION SYSTEMS},* vol. {36}, nº {2}, pp. {303-334}, {AUG} {2013}.

[92] Q. Wan, S. Huang e M. Wei, "Research on pretreatment of questions based on large-scale real questions set," *Journal of Networks,* vol. 8, nº 8, pp. 1810-1817, 2013.

[93] D. Wang e H. Zhang, "Inverse-category-frequency based supervised term weighting schemes for text categorization," *Journal of Information Science and Engineering,* vol. 29, nº 2, pp. 209-225, 2013.

[94] F. Wang, Z. Yang, Z. Li e J. Zhou, "Query Classification by Leveraging Explicit Concept Information," em *Advanced Data Mining and Applications: 12th International Conference, ADMA 2016, Gold Coast, QLD, Australia, December 12-15, 2016, Proceedings 12*, 2016.

[95] P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu e H. Hao, "Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification," *Neurocomputing,* vol. 174, pp. 806-814, 2016.

[96] P. Wang, J. Xu, B. Xu, C.-L. Liu, H. Zhang, F. Wang e H. Hao, "Semantic Clustering and Convolutional Neural Network for Short Text Categorization.," em *ACL (2)*, 2015.

[97] R. Yadav e M. Mishra, "QUESTION CLASSIFICATION FOR QUESTION ANSWERING SYSTEM USING BACK PROPAGATION FEED FORWARD ARTIFICIAL NEURAL NETWORK (BPFFBNN) APPROACH," 2013.

[98] A. A. Yahya, A. Osman e M. S. El-Bashir, "Rocchio algorithm-based particle initialization mechanism for effective \{PSO\} classification of high dimensional data," *Swarm and Evolutionary Computation* , pp. - , 2016.

[99] Y. Yoshikawa, T. Iwata e H. Sawada, "Latent support measure machines for bag-of-words data classification," em *Advances in Neural Information Processing Systems*, 2014.

[100] L. Zhang, Y. Li, Y. Xu, D. Tjondronegoro e C. Sun, "Centroid Training to achieve effective text classification," em *2014 International Conference on Data Science and Advanced Analytics (DSAA)*, 2014.

[101] D. Zhang e W. S. Lee, "Question classification using support vector machines," em *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003.

[102] R. Zhang, H. Lee e D. Radev, "Dependency sensitive convolutional neural networks for modeling sentences and documents," *arXiv preprint arXiv:1611.02361,* 2016.

[103] Y. Zhang, S. Roller e B. Wallace, "MGNC-CNN: A simple approach to exploiting multiple word embeddings for sentence classification," *arXiv preprint arXiv:1603.00968,* 2016.

[104] Z. Zhang, H. Lin, P. Li, H. Wang e D. Lu, "Improving semi-supervised text classification by using Wikipedia knowledge," em *International Conference on Web-Age Information Management*, 2013.

[105] C. Zhou, C. Sun, Z. Liu e F. Lau, "A C-LSTM neural network for text classification," *arXiv preprint arXiv:1511.08630,* 2015.

[106] B. S. Bloom, "The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring," *Educational researcher,* vol. 13, nº 6, pp. 4-16, 1984.

Valtemir A. Silva started working ~~toward the~~on his PhD degree in ~~computer~~ Computer ~~science~~ Science in July 2014 at the ICMC – University of São Paulo (USP). He received the master's degree in ~~production~~ Industrial ~~engineering~~ Engineering in 2006 ~~at~~ from the EESC – USP. ~~In the context of the PhD, he is~~His areas of interest are ~~concerned with~~ software engineering and computational intelligence. His research investigates problems regarding text classification, question classification, adap~~a~~tive learning and relationship between human skills and learning ~~requisites~~ requirements for evaluations.

Ig Ibert Bittencourt is an Associate Professor at Federal University of Alagoas (Brazil), the former president of the Special Committee on Computers and Education from Brazilian Computer Society (leading around 2500 researchers) and editor of IEEE Transactions on Learning Technologies. Prof. Ig Bittencourt co-founded an awarded company called MeuTutor (more than 50 thousand students already used) and he stands out from his peers by creating one of the most innovative companies in the field of educational technology in Brazil. He believes in innovative social entrepreneurship as a model for promoting a sustainable economic and social development to mankind.

José C. Maldonado received his bachelor's degree in Electrical Engineering from the University of São Paulo, São Carlos/Brazil, in 1978, his master's degree in Spatial Engineering and Technology from the National Institute For Space Research, São José dos Campos/Brazil, in 1983, and his PhD in Electrical Engineering - Computing and Automation from the University of Campinas, Campinas/Brazil, in 1991. He was a postdoctoral researcher at the Purdue University, West Lafayette, Indiana/United States, in 1996, and since 1997 has been a Professor at the University of São Paulo, São Carlos/Brazil.