Check for updates

EXTERNAL SCIENTIFIC REPORT

APPROVED: 15 May 2018

doi:10.2903/sp.efsa.2018.EN-1427

Machine learning techniques for the automation of literature reviews and systematic reviews in EFSA

Stijn Jaspers[1], Ewoud De Troyer[1], Marc Aerts[1]

[1]Interuniversity Institute for Biostatistics and statistical Bioinformatics, UHASSELT, Diepenbeek, Belgium

Abstract

This Report presents the results from EFSA project RC/EFSA/AMU/2016/01 related to the implementation of machine learning techniques for literature reviews and systematic reviews in EFSA. An overview of the different steps of a systematic review is provided, along with possible ways for automation. Although it was found that most steps could benefit from automation, it was also observed that some steps require more sophisticated methods than those encompassed within the machine learning framework. Availability of data and methodology allowed for the development of an automatic screening tool based on several machine learning techniques. The developed shiny R application can be used for the screening of abstracts and full texts. Properties of machine learning techniques are discussed in this Report together with their most important advantages and disadvantages. The latter discussion includes both general properties, as well as context-specific properties based on their performance in three case studies. Although creating a universal automatic data extraction tool was considered to be infeasible in this stage, this step of the systematic review was addressed to allow the reviewer to scan the uploaded pdf files for certain words or string of words. Based on observations from the performed case studies, recommendations were made regarding which methods are preferred in specific situations. More explicitly, a discussion is made about the performance of the classifiers with respect to the magnitude of the pool of papers to be screened as well as to the amount of imbalance, referring to the proportion of relevant and irrelevant papers. Finally, it was concluded that the results presented in this report provide proof that the developed shiny application could be efficiently used in combination with other software such as DistillerSR.

© European Food Safety Authority, 2018

Key words: Systematic Reviews, Machine Learning, screening, data extraction, Sensitivity, Specificity

Question number: EFSA-Q-2016-00294

Correspondence: AMU@efsa.europa.eu

Disclaimer: The present document has been produced and adopted by the bodies identified above as author(s). This task has been carried out exclusively by the author(s) in the context of a contract between the European Food Safety Authority and the author(s), awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the authors.

Acknowledgements: The authors would like to thank the EFSA staff members Ana Garcia, José Cortiñas Abrahantes, Federica Barrucci, Laura Martino, Irene Muñoz Guajardo, Ermanno Cavalli, Carsten Behring, Chiara Bianchi, Raquel Garcia Matas, Sandra Correia, Maria Rosaria Mannino, Blaize Abuntori, Filippo Bergeretti, Benjamin Taiwo Abegunde, and Didier Verloo

Suggested citation: Jaspers S, De Troyer E, Aerts M, 2018. Machine learning techniques for the automation of literature reviews and systematic reviews in EFSA. EFSA supporting publication 2018:EN-1427. 83 pp doi:10.2903/sp.efsa.2018.EN-1427

ISSN: 2397-8325

© European Food Safety Authority, 2018

Reproduction is authorised provided the source is acknowledged.

Summary

This report presents the results from objectives 1 to 5 of EFSA project RC/EFSA/AMU/2016/01 related to the use of machine learning techniques for literature reviews and systematic reviews in EFSA.

As such, an overview of the different steps of a systematic review (as introduced in Tsafnat et al., 2014) is given, along with possible ways for automation. Automation through the use of machine learning techniques was proposed for three different steps: the screening of abstracts, data extraction and critical appraisal. The implementation of the machine learning techniques for the screening of abstracts and full texts was considered to be feasible. The basic properties of these machine learning techniques were discussed, together with their most important pros and cons.

Results of three cases studies are presented. The majority of the results are related to machine learning techniques used to automate the abstract screening step of a systematic review. The data of these case studies were obtained from different units within EFSA and are referred to as the Isoflavones (published EFSA Opinion), QPS and ERIS case studies. For all datasets, support vector machines, gradient boosting machines, neural networks and random forests classifiers were constructed using either the term-document matrix or the topics approach as feature space. All case studies showed severe class imbalance. Hence, in addition to training the classifiers on the original data, the SMOTE and ROSE sampling adjustments were applied as well.

In addition to a discussion of the performance of the individual classifiers, ensembles of the constructed classifiers were created as well and their performance discussed. It was noted that these latter ensembles often had a good performance in terms of identifying relevant abstract, while reducing the workload for the reviewer. Nevertheless, the optimal ensembles are often hard to find and individual classifiers could therefore be considered to be an adequate alternative. Especially the random forests and neural networks, trained using topics with the adjustment for class imbalance, performed similar to the ensemble of individual classifiers.

Next to abstract screening, the methodology was also applied to the task of full text screening. However, it was observed that the amount of correctly identified relevant abstracts was smaller for the full text screening as compared to the abstract screening, but the amount of correctly identified irrelevant articles was higher. This inability to identify relevant papers was mainly due to the fact of the limited amount of available training data and a more in-depth study is recommended for the future.

Although creating a universal automatic data extraction tool was considered to be infeasible at this stage, this step of the systematic review was addressed in a more basic manner. More specifically, the R shiny tool that was constructed to aid in the screening for relevance also allows for searching the pool of full texts for specific, user-defined words and to show the context in which these words appear. In this way, the user can more efficiently retrieve specific data elements from the papers. A detailed explanation of data extraction is given in an accompanying tutorial of the shiny application.

In summary, the authors believe that the results presented in this paper provide proof that the developed shiny application could be efficiently used in combination with the DistillerSR tool from EFSA, where the latter can create the input files in the correct format and the shiny application could be used to reduce the manual workload for the reviewers in the steps of screening abstracts and full texts. Also searching for important key words in the pool of relevant articles can be done in a simple, yet fast manner using the application.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Table of contents

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

1. Introduction

1.1. Background and Terms of Reference as provided by the requestor

This contract was awarded by EFSA to: Universiteit Hasselt

Contractor: Universiteit Hasselt

Contract title: Assistance to the Assessment and Methodological support Unit for the provision of services to EFSA on the use of machine learning techniques for literature reviews and systematic reviews in EFSA

Contract number: RC/EFSA/AMU/2016/01

In EFSA context, the production of a scientific assessment might include the performance of a literature review or a Systematic Review (SR). A systematic review is a structured process that includes the selection, appraisal and synthesis of all the relevant evidence found in relation to a specific research question or sub-question. This process is increasingly used by the scientific community and considered a cornerstone of evidence-based research. At the same time, it poses challenges due to the exponential growth in evidence production and the consequent high demand in terms of time and resources (Tsafnat et al., 2014). Therefore, the automation of some of the tasks to be performed in a literature review or a systematic review is highly desirable since it will assist in making these processes more feasible particularly when performed on a routine basis.

The definition of Machine Learning (ML) in this procurement follows the definition given by Mitchell (1997): "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E". The emphasis of this procurement procedure should be placed on automatic methods that can be reproducible in EFSA routinely.

At the present time, some methodologies falling under the broad category of Machine Learning Techniques (MLTs) have started to be used in specific steps of the SR process and their performance tested with respect to the golden standard (e.g. expert reviewers performing the screening of papers). Although the results corresponding to some of these tests are encouraging, further research is needed to assess the applicability of these methodologies in literature reviews and systematic reviews more routinely.

In 2015, EFSA started an outsourced project on the use of MLT technologies focusing on MLT use for the statistical aspects of EFSA scientific assessments (P-AMU-10: EFSA-Q-2014-00467).

In this report, the purpose is to focus on MLT techniques that could be used in EFSA during the production of literature reviews and/or systematic reviews in order to harmonise and to streamline the process saving time and resources. As such, this report links with two other relevant activities in EFSA: R Services for EU projects (R4EU) and Critical Appraisal Tools (CATs) for literature reviews and systematic reviews.

In this respect, EFSA, at the forefront of scientific excellence, should i) investigate to which extent MLT techniques can be applied successfully for the production of literature reviews and systematic reviews in its scientific assessment framework; ii) assess potential benefits and challenges associated with the use of MLT techniques in EFSA and iii) pilot selected MLTs in already produced systematic reviews in order to identify strengths, weaknesses, opportunities and challenges for the routine implementation of selected MLTs in EFSA assessments.

Machine Learning Techniques are attracting a lot of interest in the field of literature reviews and systematic reviews used in scientific assessments. The implementation of MLT techniques in the generation of literature reviews and systematic reviews in EFSA in combination with specialised

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

expertise will in turn contribute to foster scientific excellence and enhance quality, credibility and trust among stakeholders and citizens.

1.2. Interpretation of the Terms of Reference

A systematic review (SR) is a highly structured process that takes several steps to be completed, often involving several reviewers working together over a large period of time. In order to decrease these high demands in terms of time and resources, several automation procedures have been discussed throughout literature with the aim of streamlining the process.

In this report, the steps identified by Tsafnat et al. (2014) are followed as a standard guideline for the production of SRs, with the addition of a critical appraisal step as this is current practice at EFSA. In addition to presenting and exploring general ways to automate the reviewing process, specific attention is addressed towards the use of Machine Learning Techniques (MLTs).

Following the definition in the terms of reference, machine learning explores the study and construction of algorithms that can learn from and make predictions on data. Hence, a MLT is a technique or an algorithm that does not strictly follow static program instructions but employs the data at hand to arrive at predictions or decisions through building a model from sample inputs. In the current application, the sample inputs are the obtained articles after performing a search related to the research question of interest. Nevertheless, it is important to note that these texts cannot be used directly as input to the algorithms. Rather, different pre-processing steps are required. In this respect, the term Natural Language Processing (NLP) is often used, which refers to the field in computer science that is concerned with the interaction between computers and human or natural language. As such, NLP is a specific component of Text Mining (TM), which performs a special kind of linguistic analysis that comes down to helping a machine read text. Once the obtained articles are transformed into a manageable input format, one application of MLTs is found in the construction of classifiers that can be employed to determine which of the retrieved articles are relevant for the study and which articles can be omitted in any further analysis.

While it is clear that NLP is required to transform the raw texts into formats that are appropriate to use as input to the MLTs, it can also be regarded as a stand-alone technique to automate certain steps of a SR, such as extracting interesting parts of the retrieved papers.

The current report introduces for each of the steps of a systematic review possible means of automation. A tabulated overview is presented to make a distinction between the general ways of automation that were identified for certain steps of the SR and the specific machine learning based automation techniques that can be used for automating other steps of the review. In this way, the report does not only focus on the specific requirement of identifying MLTs for the automation purpose, but presents a more general overview of other techniques as well. However, with the terms of reference as a background, major attention was given to the specific steps of the SR that are able to benefit from automation through the use of MLTs.

Three case studies are explored to compare the performance of the introduced MLTs for the screening of abstracts step in a SR. These methods include support vector machines (with linear, polynomial and radial kernels), gradient boosting machines, neural networks and random forests. In a later stage, full text screening was also investigated for one of these case studies to check whether including the full text in the screening process would lead to better performance results. Finally, it was argued in the same interim report that creating a universal automatic data extraction tool is not feasible. Nevertheless, this step of the systematic review was addressed as well, albeit in a more basic manner. More specifically, the R shiny application that was constructed to aid in the screening for relevance also allows for searching the pool of full texts for specific, user-defined words and to show the context in which these words appear. In this way, the user can more efficiently retrieve specific data elements from the papers. With respect to the critical appraisal step, general possibilities for automation are described, but not implemented. The main reason for this was the need for more in-depth data, in which information was present not only on the article level, but also on the within-article level. The within-article level refers to the assignment of a relevance indicator to specific sentences that tells the

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

machine whether or not the specific sentence indicates a possible bias in the article. Moreover, the proposed methodologies for this latter step of the reviewing process are only at a preliminary stage and further fundamental research is required. Therefore, after several efforts and discussions with EFSA, the implementation of an automatic critical appraisal stage was considered to be out of the scope of the project.

Following a short discussion during the interim meetings, it was argued that the developed R shiny application could be integrated with the DistillerSR tool from EFSA, where the latter can be used to create the input files in the correct format and the application could be addressed to reduce the manual workload for the reviewers in the steps of screening the abstracts and full texts. No other links for integration could be identified and focus was addressed to the discussion of the case study results (i.e. objective 4).

1.3. Additional information

The general objective in this project was to obtain better insights into selected Machine Learning Techniques for the routine generation of literature reviews and systematic reviews in EFSA. In order to achieve this general objective, focus was addressed to achieving the following specific objectives:

Objective 1: To identify MLTs that can be potentially applied in the production of literature reviews and systematic reviews in EFSA routinely. In order to identify MLTs that could be used in EFSA, additional information was gathered and literature searches were performed. Furthermore, it was investigated to which extent and in which steps of the process (all or several steps of literature searches and/or systematic reviews) identified ML techniques can be applied. Following both the EFSA guidance and Tsafnat et al. (2014), all the steps of literature searches and/or systematic reviews were first identified and described. Next, for each of these steps, it was assessed whether the task/step of the literature search and/or systematic review could be potentially undertaken by the use of MLTs (for example if MLTs could be used for the search of relevant studies and/or for the critical appraisal of studies/scientific outputs, etc.). Related to this specific objective, a lot of work has been done at EFSA before. Especially the work done by Ru et al. (2016) proved to be very useful.

Objective 2: To evaluate potential benefits and challenges associated with the use of identified MLTs for the production of literature reviews and systematic reviews (or parts of the process) in EFSA. For each identified MLT, the main characteristics should be described as well as the condition of application; the availability of the technique in statistical software used in EFSA should be assessed as well as its validity and reliability when used for specific steps of literature reviews and systematic reviews in EFSA.

Objective 3: To identify specific links or areas of integration of resources with related EFSA on-going activities such as, for example, the investigation into data mining techniques used for the identification of emerging risks. In addition, objective 3 is also dedicated to investigate the integration of already existing tools with new identified promising MLTs in diverse EFSA scientific areas.

Objective 4: To test or pilot the application of the most reliable MLTs on already produced systematic reviews. In this respect, several case studies were performed, in which we used several of the identified MLTs. In the case of machine learning, the contractor should assess the human resources needed or employed in order to instruct and/or test the machine if/when applicable.

Objective 5: To produce a final report including SWOT analyses, sensitivity and specificity levels and recommendations regarding the use of tested methodologies to assist in the generation of literature reviews and systematic reviews in EFSA and to provide EFSA with a fully documented list of MLTs, R codes and algorithms that can be readily used in EFSA routinely.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

2. Data and Methodologies

Several machine learning techniques (MLTs) have been considered in order to automate specific steps of a systematic review. The performance of the MLTs in the screening part of a systematic review has been assessed. In order to do this, three EFSA case studies were investigated, i.e. the Isoflavones, QPS and ERIS case studies. Some basic information on these studies can be found in Section 2.1. In Section 2.2, the different steps of a SR are discussed together with the possibilities for automation, either through MLTs or using more general techniques.

2.1. Data

In 2015, EFSA prepared a scientific opinion on the possible association between the intake of isoflavones from food supplements and harmful effects on mammary gland, uterus and thyroid in peri- and post-menopausal women (EFSA ANS Panel, 2015). In this perspective, a systematic review was performed. For the purpose of this project, a csv file containing title and abstract information on 6867 articles was provided by EFSA and used to test the MLTs that were selected for the abstract screening task. For this large pool of abstracts, an indicator of relevance was assigned in the following way: an abstract was considered to be relevant in case EFSA reviewers had indicated that the full text should be screened. In this way, 383 articles were found to be relevant (5.6%), while the other 6484 (94.4%) articles were considered to be irrelevant to the systematic review. It is noted that the majority of the articles are irrelevant and that the classification is highly imbalanced. This dataset will constitute the first case study of interest.

One of the tasks of EFSA is to assess the safety of a broad range of biological agents in the context of notification for market authorisation as sources of food and feed additives, food enzymes and plant protection products. Therefore, the Qualified Presumption of Safety (QPS) assessment was developed to provide a harmonised generic pre-assessment to support safety risk assessments performed by EFSA's scientific Panels. One of their tasks is to assess microorganisms including bacteria, yeasts and viruses used for plant protection purposes following an Extensive Literature Search strategy. The second case study of interest in this report uses Bacillus data obtained from QPS. More precisely, information on the title and abstract of 4091 articles was used. Reviewers of EFSA indicated whether or not an article could be considered to be relevant. In this setting, the reviewers had three choices i.e. 'Yes', 'No' or 'Maybe'. In case the answer is 'Maybe', the article should go to the full text screening phase, so when training the MLTs for automatic abstract screening, these articles were considered to be relevant. In this way, the pool of abstracts for this case study consists of 161 relevant abstracts (3.9%) and 3930 irrelevant abstracts (96.1%). Again here, there is a pronounced class imbalance.

The third and final case study results from SCER Unit in EFSA and will be referred to as the ERIS case study. In this case study, 668 articles were included, among which 556 were considered to be irrelevant (83.2%) while 112 were labelled to be relevant (16.8%). It can be noted that this dataset is also imbalanced, but less pronounced compared to the former two, with a minority class consisting of 16.8% of all abstracts. Nevertheless, the amount of data available here is much smaller compared to the former two case studies. Since this lack of data availability might have a negative impact on the training process, MLTs were trained here as well to automatically determine the relevance of the abstracts and the performance measures are discussed.

2.2. Methodologies

In order to arrive at the final overview of possible ways to automate the process of performing systematic reviews, several searches were performed. To get an initial idea about which steps could possibly benefit from automation, a basic internet search was conducted using the Google search engine. The employed search query was very general, i.e. "automating steps of systematic reviews". Since the goal of this initial search was to gain basic insights, only the first three pages of results were considered (i.e. 33 items). Among them, four review papers were found to be related to the question of interest. Among these four, three addressed the automation of the entire SR process (i.e., Hamad

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

and Salim, 2014; Tsafnat et al., 2014; Tsertsvadze et al., 2015). The fourth article, being Jonnalagadda et al. (2015) specifically focused on the data extraction step of the review. Therefore, mainly the three general overviews were regarded this stage.

It was noted that the conclusions in Hamad and Salim (2014) were only very basic, pinpointing MLTs only for one particular stage of the SR. A better, more extensive overview was provided by Tsafnat et al. (2014), who paid attention to all distinct steps of the SR. In addition, when focusing on automation through MLTs in the review by Tsertsvadze et al. (2015), it was noted that they mainly referred to the overview by Tsafnat et al. (2014) for a general idea of automation. In terms of specific steps of the reviewing process, referrals were made to Jonnalagadda et al. (2015) for data extraction and O'Mara-Eves et al. (2014) for the screening of abstracts. Therefore, in the following sections, these latter three papers constitute the main building blocks and were used as a basis to perform additional searches.

2.2.1. Different Stages of a Systematic Review

Conducting a systematic review is often a time-intensive process that may require a large amount of resources. In order to conduct a good SR, it is advisable to follow a series of steps related to specific tasks. An overview of the different steps, tasks and stages of a systematic review can be obtained in Table 1: . This table was modified from Tsafnat et al. (2014) through the extension with a critical appraisal step (step 12), thereby following the EFSA guidance.

Table 1: Different steps, tasks and stages of a systematic review as modified from Tsafnat et al. (2014)

Step/Task	Description	Stage
1. Formulate review question	Decide on the research question of the review	Preparation
2. Find previous systematic reviews	Search for SR that answers the same question, (part of scoping the literature in EFSA guidance)	Preparation
3. Write the protocol	Provide an objective, reproducible, sound methodology for peer review	Write up
4. Devise search strategy	Decide on databases and keywords to find all relevant trials	Preparation
5. Search	Aim to find all relevant citations even if many irrelevant ones are included	Retrieval
6. De-duplicate	Remove identical citations	Retrieval
7. Screen abstracts	Based on titles and abstracts, remove definitely irrelevant trials	Screening
8. Obtain full text	Download or request copies from authors	Retrieval
9. Screen full text	Exclude irrelevant trials	Screening
10. Snowball	Follow citations from included trials to find additional ones	Retrieval
11. Extract data	Extract relevant information (either quantitative or qualitative) to help with the synthesis and conclusions	Synthesis
12. Critical appraisal	Assessing the risk of bias in the included studies	Critical Appraisal/ Synthesis

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Step/Task	Description	Stage
13. Synthesize data	Convert extracted data to a common representation considering the results from the critical appraisal (if /when applicable)	Synthesis
14. Re-check literature	Repeat search to find new literature published since the initial search	Retrieval
15. Meta analyse	Statistically combine the result from all included trials	Synthesis
16. Write up review	Produce and publish final report	Write up

It is observed that the entire process of performing a SR is composed of 5 different stages. The focus of the initial preparation stage is to clearly formulate the research question (step 1) and devise a search strategy (step 4). The possible automation of this stage is discussed in Section 2.2.1.1. These initial conclusions should be summarized into a protocol (step 3), where the reviewer provides an objective, reproducible, sound methodology for peer review (first part of the write-up stage). Consecutively, in the retrieval stage, the aim is to retrieve all relevant citations (step 5) to answer the research question of interest, even if many irrelevant citations are included. Often, only the titles and abstracts are retrieved at this point, which are used to identify and remove the irrelevant citations in the first part of the screening stage (step 7). Next, the full texts of the possible relevant citations are retrieved (step 8) and again screened (step 9) for their actual relevance. From this, it is clear the retrieval and screening stage are closely intertwined. Hence, the possible automation with MLTs of these stages will be discussed together in Section 2.2.1.2. Once all relevant citations and their corresponding full texts are obtained, the reviewer needs to extract the relevant information (step 11) from each paper, perform critical appraisal of papers/studies if applicable (step 12) and combine the results in the synthesis stage (step 13). Automation of this stage is discussed in Section 2.2.1.3. Of course, since a SR is a time-consuming process, the reviewer should perform an additional search for new relevant papers around this stage (step 14). Once this is performed and possible new results are appraised (if applicable) and synthesised as well (step 15), the final report can be written and published, thereby completing the write-up stage (step 16). Since the writing of a report is an important task that highly depends on the specific domain and interests of the dedicated community, the automation of this stage will not be discussed.

In order to assess which steps of the systematic review could be automated, existing systematic reviews in this field were queried first (thereby following step 2 from Table 1:). In addition to the paper by Tsafnat et al. (2014), four SRs addressing the concept of automation were found. Hamad and Salim (2014) concluded that much work had been done to automate the study selection process, but they found no evidence about automation of the planning and reporting process (i.e. preparation, retrieval and synthesis stages in Table 1:). Nevertheless, these stages can benefit as well from automations and possible approaches are listed in the following subsections. In addition, O' Mara-Eves et al. (2014) identified papers that use MLTs in the screening of abstracts stage (i.e. step 7 in 0). In this respect, they identified 44 relevant papers, addressing a variety of possible MLTs. With respect to these 44 different studies that were presented in O'Mara-Eves et al. (2014), Olorisade et al. (2016) present a critical analysis comparing their performance. Finally, Jonnalagadda et al. (2015) performed a SR to identify articles dealing with the automation of data extraction. While they identified 26 published reports describing automatic extraction of at least one of the more than 50 potential data elements used in systematic reviews, they concluded that no unified framework for automatic information extraction has yet been developed. Throughout the next subsections, the reflections made in the mentioned systematic reviews are presented alongside some additional thoughts on possible automations for all stages of the SR.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

2.2.1.1 Preparation Stage of a Systematic Review

In the preparation stage, three tasks are central, i.e. formulating the review question/s, finding previous SRs and devising the search strategy. First of all, it is of great importance to formulate the research question/s in a clear and precise way. Only then, a transparent and reproducible review can be performed. Next, the workload can be greatly reduced by querying for systematic reviews that address the same research question/s. Indeed, since performing a SR is often time and resource consuming, one should first determine whether a good quality review of the topic of interest does not already exist. In case one does exist, one should assess the quality of the review and whether it is in need of updating. In many cases, even if an out-of-date SR has been identified, updating it according to an established protocol is preferable to conducting a new review (Tsafnat et al., 2014). Different databases should be queried before a new research is started. Of course, the list of databases to be queried highly depends on the topic of interest. For example, the most commonly used medical databases include PubMed, Scopus, Web of Science and Google Scholar. The latter three databases also provide information on other fields, such as science and technology, social sciences and arts and humanities.

At the end of the preparation stage, the researcher should have gathered enough preliminary information in order to write up the protocol. In this respect, several templates can be employed to assist the researcher in performing the appropriate steps (Higgins et al., 2011 and Cochrane Collaboration, 2013). For example, Cochrane's Review Manager uses a protocol template that provides standard fields to remind the reviewer to cover all required aspects of specific protocols. Of course, different research fields might have different requirements, so the template suggested here is not a universal tool. Ongoing research is performed to improve existing templates (Higgins et al., 2011; Tsafnat et al., 2014).

To round up the preparation stage, a search strategy should be presented. More specifically, the reviewer should decide on what keywords will be used, which databases will be queried and how certain citations will be tracked. Regarding the latter, Kuper et al. (2006) describe the potential benefits of using the Science Citation Index in combination with a PubMed search. The authors concluded that not using citation tracking in a systematic review of observational studies may result in bias. Regarding the choice of keywords to be employed, natural language processing might be interesting to automatically understand a research question and its context. In addition, next to querying existing databases, one should also aim to identify grey literature on the web and institutional repositories. By also including these studies, the final SR will be less prone to bias.

2.2.1.2 Retrieval and Screening Stage of a Systematic Review

After the protocol and search strategies have been finalized, the next aim in the systematic review is to collect all relevant citations regarding the research question of interest. A lot of databases already incorporate so called automatic query expansion algorithms to optimize the search. These algorithms include, amongst others, synonym expansion (i.e. automatically adding synonymous key words to the query), word sense disambiguation (i.e. understanding keywords in the context of the search and replacing them with more sensible synonyms) and correct spelling mistakes.

In a related EFSA project on the identification of possible MLTs, automated search procedures were constructed for the specific task at hand (Ru et al., 2016). For example, articles that are available in the online repository arXiv can be searched and retrieved using the aRxiv R package. Similarly, the RISmed R package provides the functionalities for querying the PubMed database. R packages related to a selection other databases exist as well, although often in an early development stage (see Table 3:). In case no R packages exist, either manual retrieval or customized R scripts (to interface the resource web page) are required. Therefore, obtaining a unified framework for automated searches is very difficult.

Often, performing searches in different databases leads to a large collection of citations, among which there can be some duplicate references. Identifying and removing duplicate records (step 6) can be performed automatically using specific citation managers. Kwon et al. (2015) compared de-duplicating

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

in two different database platforms (Ovid and EBSCO) with de-duplicating in three citation managers (RefWorks, EndNote, Mendeley). It was noted that there is not a clear consensus on which method performs best. When performing a SR, reviewers want to maintain the highest possible recall in retrieving citations. Among the three citation managers, EndNote performed the worst, having both the highest number of false positives and false negatives. The authors noted that these results were in line with the findings of Qi et al. (2013) and Rathbone et al. (2015). More specifically, these authors argued that the automatic deduplication option in EndNote is not fully adequate and should be supplemented by a manual search. Rathbone et al. (2015) developed the Systematic Review Assistant-Deduplication Module (SRA-DM), which was found to be superior to the EndNote deduplication method in terms of sensitivity and specificity. Finally, a similar study was performed by Bramer et al. (2014), who identified RefWorks to be the least effective citation manager. Their algorithm, which relies on the EndNote software, was found to outperform the algorithm developed by Qi et al. (2013) and other citation managers, including Mendeley, Jabref, Refworks, Zotero, Paperpile, EndNote (standard) and refman. Hence, when the standard practice is to use the EndNote software for deduplication, it might be advisable to use the Bramer method for better performance. The implementation of this method is clearly described in Bramer et al. (2016).

Of course, some issues, such as "studification", might still remain. The term "studification" refers to the situation where the same study has multiple reports with possibly a different list of author names, a different title or appeared in different journals. These references should often all be cited, but should only be used once in the meta-analysis (in order to avoid biased results).

Once the reviewers have obtained their initial pool of possibly relevant citations, one should first screen (step 7) the corresponding titles and abstracts, in order to determine the actual relevance. In case the obtained pool consists of many citations, this step in the SR process can take a lot of time. Moreover, mistakes are easily made and a second reviewer is usually addressed to determine the relevance of the same pool of citations. Hence, particularly in this phase of the review process, automatic screening systems could be used to resolve disagreements between the two reviewers or even to replace one or both of the screeners. O'Mara-Eves et al. (2015) performed a systematic review to identify different text mining techniques which are used for study identification in systematic reviews. They identified 44 relevant papers, published between January 2006 and January 2014. Among these studies, 30 described methods that can be used to reduce the number of citations that are needed to be screened, 6 used text mining as a second screener, 7 focused on increasing the speed of screening and 12 improved the workflow through screening prioritisation. Note that these numbers do not sum up to 44 since some studies adopted more than one approach to workload reduction. Olorisade et al. (2016) re-reviewed the papers identified by O'Mara-Eves et al. (2015) and provided a more detailed overview of the different text mining methods that were used throughout. Because of the importance of having an automated screening of abstracts step, Section 2.2.2 provides further detail to this part of the systematic review. More specifically, the results found by O'Mara-Eves *et al.* (2015) and Olorisade et al. (2016) are further discussed there, together with a new search on MLTs for the years that were not covered by the previous SR (i.e. 2014-2016). Based on these literature searches, classification tools can be constructed using one or more of the identified MLTs.

Although the focus in Section 2.2.2 will be on the classification of abstracts into relevant and irrelevant groups, other workload reduction methods exist. For example, SWIFT-review provides a workbench with tools to assist the reviewer in literature prioritisation (Howard et al., 2016). Using these tools, the reviewer can determine which articles to screen first, thereby reducing the amount of time spent on irrelevant papers.

Once the abstracts and titles have been screened for their relevance, the researcher should aim to collect the full texts (step 8) associated to the relevant citations. In order to obtain the full texts, one is often faced with an entire network of links that span multiple websites. Related issues are the cumbersome subscription models or a limited archival and electronic access. Often, the queried databases already provide links that may offer additional access options. The R package 'fulltext' provides a single interface to many sources of full text, including Biomed Central, Public Library of Science, Pubmed Central, eLife and arXiv, amongst others. In addition, the 'metagear' R package

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

provides a useful function for downloading the full text PDFs. Of course, it should be noted that the download success of these PDFs is entirely conditional on the journal subscription coverage of the host institution running the R package.

After the full texts are obtained, the decision support systems that were created for abstract screening might again be used here to confirm the relevance of the citations. More specifically, instead of constructing classifiers based on the titles and abstracts only, one could also include information from the full text (step 9). To our knowledge, this kind of classification has not been done before so the outcome of this approach is unknown at this point. Nevertheless, it might be worthwhile to further explore, especially in case the pool of abstracts that are identified to be relevant is rather large. Some more advanced methods, which incorporate also tables and figures, are presented by Thomas et al. (2011) and by Rodriguez- Esteban and Iossifov (2009).

In step 10 of the SR process, the researcher should follow the citations of the selected, relevant trials to find some additional sources of information that can be of interest for the research question. This step is referred to as snowballing (Tsafnat et al, 2014). The only reference to automating this step was found to be Choong et al. (2014), who evaluate an automatic evidence retrieval system based on a modified version of ParsCit (Councill et al., 2008), where the latter is specifically developed for extracting reference strings from text documents using natural language processing (NLP) techniques. Each reference that is found from the middle to the end of the text is converted to a search engine query by removing stop words, numbers and punctuation. The final result of the query is basic citation information and often a link to the full text. These links were then extracted and followed to obtain the new texts. Note that the latter is not an automatic step, so some human input is still required. Apart from snowballing, at this step one could also search for articles that cite the relevant selected paper.

2.2.1.3 Critical Appraisal and Synthesis Stage of a Systematic Review

In the synthesis stage, the focus is on summarising the interesting results from the pool of relevant articles that were selected in the previous stage. In summary, this stage is comprised of four steps: data extraction, critical appraisal of data, data synthesis and meta-analysis (if/when appropriate).

Extracting the useful data and results from the relevant papers is one of the most challenging tasks of the SR. Therefore, a large amount of time and resources can be saved through automation of this step. According to Tsafnat et al. (2014), the approach that is currently used to automatically extract useful data consists of two stages. Initially, the amount of text to be explored is reduced by employing information-highlighting algorithms. For example, Kiritchenko et al. (2010) have developed a tool called ExaCT, which consists of an information extraction engine that searches the article for text fragments that best describe the trial characteristics, and a web browser-based user interface that allows human reviewers to assess and modify the suggested selections. In the second stage, the extracted elements are associated with variables of interest such as the treatment group or the main outcome of the trials under investigation.

The current approaches are only focused on processing one sentence at a time, making them less attractive in case of large amounts of text. In addition, more research on how to extract outcomes from large amounts of text is highly required. Due to the enormous diversity of performed studies, and related to that, the vast number of primary outcomes of interest, it is difficult to create a universal tool for extracting data from studies. In addition, due to the fact that each scientific field and each research group might have its own vocabulary, similar research subjects and trial outcomes can be described in different ways. Of course, this all adds to the complexity of creating a universal tool for data extraction.

Jonnalagadda et al. (2015) created a systematic review related to automating data extraction in SRs. They concluded that no unified information extraction framework was found that was tailored to the systematic review process. One of the major challenges in this respect is the vast amount of different data elements (>50) used in systematic reviews (Higgins and Green, 2011). All of the identified papers only dealt with a very limited number of these data elements, with a maximum of 7.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Nevertheless, while a universal tool seems out of the scope because of the abovementioned difficulties, several attempts have already been undertaken to automatically retrieve data in more specific scientific areas. Summerscales et al. (2011) present a conditional random field classifier for automatically calculating summary statistics like the absolute risk reduction and number needed to treat in clinical trials (see also Lafferty et al., 2001). Other examples include Rosario and Hearst (2005), who focused on a probabilistic graphical model for identifying treatments and diseases in sentences from medical texts and classifying their relationships. Peak et al. (2006) aimed to identify agent, patient and effect entities in sentences with specific key verbs in the conclusion section of abstracts. They employed the technique of shallow semantic parsing to achieve this aim. Another example is presented by Leaman and Gonzalez (2008), who developed BANNER, a named entity recognition system, primarily intended for biomedical text. It is a machine-learning system based on conditional random fields and contains a wide survey of the best features in recent literature on biomedical Named Entity Recognition (NER).

In addition to the conclusions in Jonnalagadda et al. (2015), an additional search was performed using the query "automatic data extraction in systematic reviews" in the Google search engine. In this way, another specific application of data extraction was found in Wallace et al. (2016), who focused on extracting PICO sentences from clinical trial reports. This method is less attractive as it requires information on previously conducted reviews (in a field related to the research question) and an available structured resource to initiate the training of algorithms. A more general information extraction algorithm was discussed by Basu et al. (2016) who employ natural language processing (NLP) and machine learning to build information extraction algorithms to identify data elements in new publications without having to go through manual annotation to build golden standards for each data type. A similar approach was provided by Bui et al. (2016). More information on these methods is provided in Section 2.2.2.

Once the available data have been obtained, the next goal is to make a synthesis. The results from the critical appraisal of studies (further explained below) may be considered at this point, for instance by only employing 'low risk of bias' studies for the data synthesis. Moreover, the data from the studies considered for the synthesis stage can be presented in a narrative or statistical manner. If studies are very heterogeneous it may be most appropriate to summarise the data narratively and not attempt a statistical summary. On the other hand, when studies are referring to similar data results, a statistical summary might be more appropriate. Indeed, although every study provides valuable information on its own, we can get a more precise and reliable estimate by combining the results of all studies into a single estimate. As such, the term meta-analysis is used to refer to an analysis of analyses (Glass, 1976). Of course, the execution of this task highly depends on the nature of the outcome of interest. Many types of meta-analysis data can be analysed using a (generalized) linear mixed model framework (Sutton et al., 2000) for which there are several R packages available. Some general R packages for performing a meta-analysis include metafor and rmeta. Furthermore, several special purpose packages for meta-analysis exist in R. Some of the packages are shortly presented below, but for a more elaborate overview, we refer to https://cran.r-project.org/web/views/MetaAnalysis.html.

- epiR: Tools for the analysis of epidemiological data. Contains functions for directly and indirectly adjusting measures of disease frequency, quantifying measures of association on the basis of single or multiple strata of count data presented in a contingency table, and computing confidence intervals around incidence risk and incidence rate estimates.
- exactmeta: Perform exact fixed effect meta-analysis for rare events data without the need of artificial continuity correction.
- MAc: This is an integrated meta-analysis package for conducting a correlational research synthesis. One of the unique features of this package is in its integration of user-friendly functions to facilitate statistical analyses at each stage in a meta-analysis with correlations.
- bspmma: Some functions for non-parametric and semi-parametric Bayesian models for random effects meta-analysis

More general packages for performing a meta-analysis include metafor and rmeta.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Finally, some attention is paid to the additional step of critically appraising the included studies in the systematic review. Indeed, as argued above, step 12 in Table 1: was added to the original steps of a SR according to Tsafnat et al. (2014) with the aim of being consistent with the EFSA guidance. Step 12 is concerned with the task of identifying the risk of bias in the selected articles. Several sources of bias exist. For example, to assess bias in the field of randomized controlled clinical trials, the Cochrane Collaboration has developed a tool that comprises 7 default domains:

- Random sequence generation, i.e. selection bias due to inadequate generation of a randomized sequence;
- Allocation concealment, i.e. selection bas due to inadequate concealment of allocations prior to assignment;
- Blinding of participants and personnel, i.e. performance bias due to knowledge of the allocated interventions by participants or personnel during the study;
- Blinding of outcome assessment, i.e. detection bias due to knowledge of the allocated interventions by outcome assessors;
- Incomplete outcome data, i.e. attrition bias due to amount, nature or handling of incomplete outcome data;
- Selective reporting, i.e. reporting bias due to selective outcome reporting
- Other sources of bias.

It has been shown that different reviewers often identify different levels of risk of bias for the same studies, most often explained by the fact that reviewers might miss key sentences (Hartling et al., 2011; Lensen *et al.*, 2014). Automating certain aspects of assessing the risk of bias can lead to reduce the time required to perform a SR and to reduce human error during the process. In spite of its importance, only a limited amount of attention has been addressed to the automation of this step. Marshall et al. (2015) and Millard et al. (2016) provide some promising approaches based on MLTs, focussing on the assessment of any of the first six domains of bias introduced above. Of course, depending on the reviewers' interest, other sources of bias might be investigated as well. The approaches are detailed upon in Section 2.2.2.

2.2.2. Automating selected steps of systematic reviews through Machine Learning Techniques

The main challenge in the application of machine learning techniques is that texts need to be presented in a format that can be used by MLTs and that provides the information that is required for the task at hand. The data of interest is the corpus, in its roughest form being just a set of texts. For further analyses, the texts in a corpus should often be annotated, a process that aims at bringing a more uniform structure to individual texts such that they can be jointly analysed. Manual annotation is very time consuming though and would not be of help in our attempt to automatically determine relevant information in texts. Therefore, if MLTs are to be used, the texts within the corpus need to be processed such that they are automatically turned into a quantitative representation.

Before models can be constructed, the rough texts should be transformed first. In an initial stage, some pre-processing takes place. This stage starts with the tokenisation of documents, which transforms a text document into smaller units which are referred to as words or terms. Typically, also certain characters are removed (e.g. non-alphabetical characters) and the conversion of characters into lower case. After the tokenisation, two other actions are often performed; i.e. stemming of words and the removal of stop words. Stemming refers to the identification of the common morphological stem of certain words (i.e. their common base or root form). For instance, the words "argue", "argued", "argues", "arguing", and "argus" all reduce to the stem "argu" and all share the same meaning. On the other hand, a stop word is a term that is often used throughout many articles of different nature and does not contribute to the meaning of a certain sentence. As such, they can be removed without affecting the connotation of the entire sentence. The R package "SnowballC" provides the functionalities required to perform these three pre-processing steps (tokenisation, word stemming and removing stop words).

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

After having pre-processed the documents, different approaches can be followed with respect to the quantification of the extracted information from the different articles. In this regard, the term feature selection is often used.

In practice, the Bag-of-words (BOW) model is mainly used as a tool of feature generation. After transforming the text into a bag of words, various measures can be calculated to characterize the text. The simplest measure is to indicate whether a term is present or not in the texts (0-1 value for the feature). Alternatively, the frequency with which terms appear in a document is also a very popular type of summary that quantifies text. Simply speaking, the more a term appears in a certain document, the higher the relevance of this term. It results in what is referred to as a Term-Document Matrix (TDM). However, term frequencies are not necessarily the best representation for the text. Common words will almost always have the highest term frequency in the text. Hence, having a high count does not necessarily mean that the corresponding word is more important. To address this problem, one of the most popular ways to "normalize" the term frequencies (TF) is to weight a term by the Inverse of Document Frequency (IDF). In addition, in the BOW approach, only the counts of words are important and the relations between the words are not taken into account. As an alternative, n-grams can be used to store some spatial information within the text. In most cases, bigrams are used which is just a combination of two consecutive words.

The advantage is clear in the sense that this simple transformation results in a numerical data representation that can be used further in various ways. As will be detailed upon below, the resulting frequencies can be used for example to determine groups of similar documents or terms and to discover which terms are indicative for detecting the documents of interest. The question of interest remains: what do frequencies of terms teach us about the task at hand? In this respect, it is important to realize that a TDM is a summary and therefore also reduces the information that was originally available in the corpus. The relations that can be explored with machine learning will in this case be restricted to relations depending solely on frequencies of terms, but not on any underlying relationships between them.

The issues with not capturing inter-word relations can also be solved by considering the Latent Dirichlet Allocation (LDA) approach (Blei et al., 2003). More specifically, LDA is an example of a probabilistic topic modelling technique, which aims at identifying topics that are present in the corpus. Hence, one can use the obtained topics instead of using the frequency of words as input variables.

Yet another alternative to the BOW approach is to use Word2Vec features (Mikolov et al., 2013), that represents each word as a vector in a form that reflects how close words occur. The closer the words tend to occur in a text the more similar the vectors will be.

These feature selection steps are very important as the machine learning techniques described in the next subsection work better on low-dimensional data. Especially in terms of time and memory savings, these steps are crucial. Different R packages can be used to derive the required feature sets. Some of them include

- “tm” package: create terms documents matrices and n-grams.
- “topicmodels” package: detect underlying topics for LDA approach
- “wordVectors” package: create the word2Vec features.

Once the features have been identified, it can be advisable to further reduce the dimension of the input space by performing term space reduction (Sebastiani, 2002). More specifically, the aim is to select and retain only part of the features that are still able to build a strong classifier. Often, simple approaches, such as retaining only features with a certain frequency across documents, suffice to perform the task with high effectiveness (Yang and Pedersen, 1997). Nevertheless, more sophisticated methods can be used as well. These include, amongst others, the odds ratio (Caropreso et al., 2001), the NGL coefficient (Ng *et al.*, 1997) and the GSS coefficient (Galavotti et al., 2000).

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

An important differentiation in statistical learning techniques is whether they are supervised or unsupervised. They differ in terms of the data that is required and in terms of what can be achieved using them.

Supervised statistical learning requires information on the true outcome or category for at least a part of the relevant data because it is learning by example. Having been trained, it can then be used to work on data for which the true outcome or category is unknown. An example of supervised learning would be to investigate whether the use of each of the terms in 100 documents would be able to correctly predict the 25 documents that are known to be of interest as opposed to the 75 that are known to be irrelevant. Once the predictions are understood, they can be applied to documents that are new.

Unsupervised statistical learning aims to externalize structure that is present in the data. An example of unsupervised learning would be to determine what documents are similar to one another based on the terms they include or do not include. Afterwards it can be established what are the dominant characteristics of each of the groups.

Both the supervised and unsupervised statistical learning techniques are of interest for systematic reviews but it should be clear that the supervised method does require information for setting up training data (i.e. a pool of papers for which the relevance is known at the time the models are constructed) and that the unsupervised method is not focused on the main task by itself (i.e. determining the relevance of the papers). Below, more details are provided on both statistical learning techniques and on specific MLTs that can be used in the screening stage and also in parts of the synthesis stage (i.e. data extraction and critical appraisal).

Using the presented case studies, a comparison was made between different feature spaces, focusing on the frequency-based TDM and the topics approach. It was observed that the TF-IDF feature space performed similar to the TDM approach. Therefore, it was not discussed in the main results section.

2.2.2.1 Supervised Learning

In general, the goal of a supervised learning problem is to use a set of input variables to predict the values of one or more output values. When interest is in determining the relevance of abstracts, the output is a 0-1 variable, taking the value 1 in case the abstract is termed relevant for inclusion in the systematic review. Similarly, in case the interest is in predicting whether a certain article is at risk of a certain bias, the output is a 0-1 variable, taking the value 1 in case the article suffers from bias.

From this point onwards, the output variable is referred to through the symbol Y. Regarding the input variables, hereafter denoted by the symbol X, several approaches can be followed. The two approaches that are frequently used are the bag-of-words (BOW) approach (using the term-document matrix as input) and the latent Dirichlet allocation (LDA) approach (using the derived topics as input). In summary, the precise goal of supervised learning is to use an appropriate feature set to create a prediction model to determine whether or not a certain article is relevant for the SR.

Olorisade et al. (2016) argued that only 35 of the 44 papers that were initially identified by O'Mara-Eves *et al*. (2015) really dealt with MLTs for the screening stage. Hence, nine papers were excluded: three because they were follow up discussions of previous results, one in which no text mining was used, three others were excluded because the methods were either not ML based or not applied in the SR context. The last paper was excluded since it mainly dealt with feature selection rather than the classification model. A summary of the remaining techniques is presented in Table 2: below.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Table 2: Machine Learning Techniques for the screening of abstracts in a systematic review identified by O'Mara-Eves et al. (2015) and discussed by Olorisade et al. (2016).

Classification Method	Number of citations	Years of publication
Support Vector Machine (SVM)	16	2006-2012;2014
EvoSVM	2	2010;2012
Naïve Bayes	7	2007;2010-2012;2014
K-Nearest Neighbour	3	2011-2012;2014
K-Means	2	2011-2012
Complement Naïve Bayes	3	2010-2012
Decision Tree	2	2007;2010
Weightily Averaged One-Dependence Estimators (WAODE)	1	2010
Neural Networks	2	2006;2012
Regression	1	2012
Ensemble	10	2006;2009-2014
Rocchio	1	2014
Distribution semantics with relevance feedback	1	2013

With the aim of identifying more papers on MLTs applied in the screening stage that were published between 2014 and 2016, a new search on Google scholar was performed. In this way, the search performed by O'Mara-Eves et al. (2015) was extended to more recent publications.

Using the query

("text mining" OR "literature mining" OR "automated") and

("citation screening" OR "article screening")

and restricting the time range between 2014 and 2017, 260 results were found. After further consideration of these papers, 31 possibly relevant articles were obtained. This initial selection was solely based on the titles of these articles, filtering articles that definitely not covered the topic automation in systematic reviews (e.g. a large number of retrieved articles dealt with the screening for particular types of cancer and were hence removed from the pool of possibly relevant abstracts). For the remaining 31 articles, the abstracts and full texts were screened and the final relevance of the papers determined.

In the end, 10 papers were ought to be relevant for the screening stage in the current overview and are discussed below. In addition, two of the remaining papers were considered to be useful in relation to the automation of data extraction and will hence also be discussed below. Also with respect to the critical appraisal stage, two papers were retained. Finally, the articles that were removed either did not describe machine learning techniques for automation or only focused on the construction of input features (such as LDA, bag of words, etc.). These features were however introduced in an earlier stage and as such, the new references did not contribute to any new information. In summary, from the 31 articles, 10 were considered to be relevant for the screening stage, 2 for the data extraction step and 2 for the critical appraisal step. It should be noted that using the same query in the Web of Science and Scopus did not provide any additional articles. A more extended search in Scopus was performed as well. Indeed, using the query

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

("text mining" OR "machine learning" OR "automating") AND ("critical appraisal" OR "abstract screening" OR "systematic review")

a total of 219 articles were obtained. Among these, 75 were retained for further investigation after inspection of the title only. Of course, there was a large amount of overlap between the new articles and those retrieved in O'Mara-Eves et al. (2015) and the formerly mentioned google scholar search. After removing these duplicates, 24 new and possibly relevant articles for the automation of specific steps of the SR remained. After screening the abstracts, 9 articles were considered to be relevant for inclusion in this report, among which 3 related to screening stage, 1 related to the critical appraisal stage and 5 for the data extraction stage.

Related to the 13 new papers related to MLTs in the screening stage, it was apparent that mainly the SVMs and the ensemble methods (such as random forest and Bayesian ensembles) received a lot of attention. For instance, Khabsa et al. (2016) created a random forests classifier using different feature spaces. In addition to working with lexical features (i.e. the BOW approach), they also included word clustering and citation features. A unified way of configuring parameters was proposed as well. Nevertheless, as argued by Saha et al. (2016), the non-lexical features such as co-citations and MeSH terms are often hard to obtain as they are not readily available. Rather, they focused on two classes of features: the uni-bigram and Word2Vec features. The first class corresponds to the BOW approach, in which not only the separate words are used as features, but also the combination of two consecutive words. The second class of features (Word2Vec) captures the semantic similarity between words similarly to LDA. Once the model is built, the features are also readily available. In contrast to LDA, there are no issues regarding the number of topics to be constructed and the features do not need to be generated per review. Employing these two feature classes, Saha et al. (2016) compared different SVMs employing different parameters and loss functions. Timsina et al. (2016) also used soft-margin polynomial SVMs as a classifier, but employed Unified Medical Language System (UMLS) for medical term extraction. These terms were consecutively employed as feature space, in addition to a second exercise where the typical BOW approach is used. The polynomial SVM was found to outperform other classifiers, including naïve Bayes, SVM with linear kernel, evoSVM and perceptron, for both feature spaces of interest. Similarly, Mo et al. (2015) employed a SVM classifier with either a linear, radial basis function or polynomial kernel on the BOW feature space as well as on topics resulting from topic modelling. They concluded that the BOW with a linear kernel SVM produced very robust results across different metrics, except for recall. Topic-based polynomial kernel SVM models provided much better recall and were hence found to be a nice alternative.

It was also observed that there seems to be a shift from supervised learning algorithms towards semi-supervised learning algorithms, which falls between supervised and unsupervised learning techniques. Indeed, it refers to methods that use a large unlabelled dataset, together with a small labelled dataset during the training stage (Wang, 2007). Liu et al. (2016) compare between 3 distinct semi-supervised learning techniques, i.e. label spreading, label propagation and semi-supervised support vector machines (S3VM). In addition, they also considered two wrapper functions, being self-training and active learning. In the former, an existing classifier is first trained with the small amount of labelled data and used afterwards to classify the unlabelled data. The most confident unlabelled points are then added to the next training set. On the other hand, active learning is a special case of semi-supervised learning. The procedure is similar to self-training. However, instead of using the labels given by the classifier, it requires a human expert to label to most confident unlabelled samples. The latter was also discussed in Wallace et al. (2010). Semi-supervised learning is also used in the text mining system FoodSIS, which was developed to improve the state of food safety in Singapore (Kate et al., 2014). The system uses several SVMs for classifying documents into a relevant and irrelevant class. In addition to the ordinary SVM, they also employed the transductive support vector machines (TSVM; Wang, 2007) and tri-class support vector machines (3C-SVM) introduced by Yang et al. (2015). The latter semi-supervised algorithms were found to outperform the supervised naïve Bayes classifier.

An alternative to MLTs was presented by Ji et al. (2015), who developed a network approach based on MEDLINE specific elements. In this way, their network can be used to discover relationships

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

between documents and use these relationships to facilitate a recommendation process. Similarly, Sellak et al. (2015) did not consider MLTs but proposed an alternative approach based on semantic rule-based classifiers. The approach involved a hybrid feature selection method within a Class Association Rules (CARs) algorithm. Since the focus in this report is on the application of MLTs, these methods will not be discussed in more detail.

New feature sets were also identified in this new search. Hashimoto et al. (2016) introduces a new topic detection method that induces an informative representation of studies with the aim of improving the performance of the underlying MLT. It uses a neural network-based vector space model to capture semantic similarities between documents. First, documents are represented within the feature space and clustered into a predefined number of clusters. The centroids of these clusters are treated as latent topics. Consecutively, each document is represented as a mixture of latent topics. Hence, this new approach can be employed as an alternative to LDA topic modelling. Similarly, Yu et al. (2016) investigated SVMs, logistic regression and decision trees on a new feature space that exploits the correspondence between topics generated using LDA and MeSH terms (referred to as TopicalMeSH). In general, SVM had the best performance, while the decision trees performed worst. As mentioned above, MeSH terms are specific to the PubMed database and are therefore not well-suited for creating a general classification tool.

The application of the approaches introduced in Table 2: is not restricted to the screening stage. Indeed, Marshall et al. (2015) use soft-margin SVM to construct a classifier to determine whether an article is prone to one or more sources of bias across clinically important areas. They developed the tool RobotReviewer, which assigns low, high or unclear risk of bias rating to a specific domain of bias and identifies fragments of text supporting these judgements. Underlying this tool, they use the BOW input features and construct a SVM classifier that is able to identify possible biases in the articles. Another SVM classifier is constructed to determine if a certain sentence in the article was used for the identification of that specific source of bias. Finally, they also present a hybrid approach that combines the output of both classifiers to construct a more powerful decision tool for the identification of bias in articles. In this respect, it is important to note that their approach was specifically developed for the critical appraisal of articles in systematic reviews of randomized controlled trials. The labelled data they required were obtained from previous studies collected in the Cochrane database. In this way, the burden of the need to create a labelled dataset was circumvented. As such, the term distant supervision is often used, as opposed to supervised learning, since the employed labelled data are not directly related to the current systematic review. Nevertheless, in case such a remote database of indirectly related articles is at hand, supervised learning can also be used. Since this requires the reviewers to manually create a labelled dataset, this is of course more time-consuming. In a similar fashion, Millard et al. (2016) use logistic regression models instead of SVMs to construct similar classifiers for the identification of several sources of bias. Lin et al. (2011) proposed a classification tree to determine the level of evidence provided by medical articles, but does not focus on automatic critical appraisal as such.

Finally, machine learning techniques can also be used with the aim of automating data extraction. For example, PICO detection in abstracts by means of SVM-based machine learning is discussed in Boudin et al. (2010a, 2010b). Likewise, Robinson (2012) applies naïve Bayes, multinomial naïve Bayes, SVM, logistic regression and random forests to the task of identifying abstracts with patient oriented outcomes based on reliable evidence. Note that both approaches deal with a very specific task, i.e. to identify papers containing a certain outcome of interest. After the classifiers have identified these papers, the reviewer can manually search the articles for the exact information that is required. As such, these approaches are not general tools that provide a unified framework for data extraction. On the other hand, the study by Basu et al. (2016), mentioned above, provides a more general approach. Their system is developed in two stages. In the start, it uses information that is contained in existing SRs to identify sentences in the included references that contain specific data elements using a modified Jaccard similarity measure. These sentences can then be used as labelled data and a support vector machine classifier is trained on these labelled data to extract data elements of interest from a new article. At this point, the system is restricted to biomedical research papers only, but the authors intend to extend it to other types of SRs. A similar approach is followed by Bui et al. (2016), except

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

that they create summaries of the texts instead, which can consecutively be used by the reviewers to draw conclusions. Note also here that an existing, related SR is required to train the classifier. From the remaining articles related to the automation of data extraction, it became apparent that MLTs alone are not a sufficient tool. Rather, there should be a symbiosis between machine learning, statistical techniques and, especially, natural language processing which includes the extraction of lexical knowledge, lexical and structural disambiguation (e.g., part of speech tagging, word sense disambiguation), grammatical inference, and robust parsing (Mishra et al., 2014).

The following sections provide more detail on the methods in Table 2: . Usually, the classifier is trained on part of the dataset, which is referred to as the training set. Once the classifier is obtained, its performance can be evaluated on the remaining part of the dataset, referred to as the test set. In addition, the following sections do also present several R packages that can be used for the implementation of the different methods. Note that the list of R packages is not extensive, but gives an indication of the packages that are most often used. During the actual implementation of the methods in objectives 3 and 4, a more thorough discussion of the possibilities of the distinct packages can be given.

2.2.2.1.1 Support Vector Machines

Support Vector Machines were developed by Cortes and Vapnik (1995) with the aim of binary classification. The basic idea of the approach is to determine the optimal separating hyperplane between the two classes of interest. This is performed by maximizing the margin between the classes' closest points. In this respect, the points that are located on the boundaries are called support vectors, and the middle of the margin is the optimal separating hyperplane. Data points on the wrong side of the discriminant margin are weighted down to reduce their influence. In case one cannot determine a linear separator, data points are projected into an (usually) higher-dimensional space where the data points effectively become linearly separable. This projection is obtained through kernel techniques (e.g. polynomial or radial basis function kernels).

In order to allow some points on the wrong side of the margin, some slack variables $\xi = (\xi_1, \ldots, \xi_N)$ are introduced. Formally, the support vector classifier is then defined by:

$$\min\|\beta\| \text{ subject to } y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, \forall i,$$

with the additional constraint that $\xi_i \geq 0$ and $\sum \xi_i \leq C$, where C is a constant. Using Lagrange multipliers, we can solve the equation for β.

The software that can be used to apply these methods: svm() function of R package e1071, ksvm() function of R package kernlab.

An evolutionary algorithm to solve the dual optimization problem of an SVM might be employed as well. This implementation is also capable of learning with kernel functions which are not positive semi-definite and can also be used for multi-objective learning which makes the selection of certain parameters in the SVM unnecessary before learning. As noted from Table 2: , this EvoSVM has only received little attention (2 citations). This might be due to the fact that it is not implemented in R. (This method is implemented in RapidMiner)

2.2.2.1.2 Naïve Bayes

The Naïve Bayes algorithm is a classification technique that is based on Bayes' Theorem, with an assumption of independence among the predictor variables. Simply speaking, a NB classifier assumes that the presence of a particular feature class is unrelated to the presence of any other feature. It is an easy-to-build model that is particularly useful for large datasets. The probability to belong to a certain class (c), given some input values or predictors (X = $(x_1, \ldots, x_n)$) is given by the following formula:

$$P(c|X) = \frac{P(X|c)P(c)}{P(x)},$$

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

where P(c|X) is the posterior probability of class (c, target) given predictor (X, attributes), P(c) is the prior probability to belong to a certain class, P(X|c) is the likelihood which is the probability of predictor given class and P(X) is the prior probability of predictor. An observation is classified to the class c for which the posterior probability is that largest, i.e. the Bayesian classifier is defined as

$$\arg\max_{c \in C} P(c|X) = \arg\max_{c \in C} P(c) \prod_{i=1}^{n} P(x_i|c)$$
$$= \arg\max_{c \in C} \left[\log p(\vec{\theta}_c) - \sum_i f_i \log \theta_{ci} \right],$$

where $\vec{\theta}_c = (\theta_{c1}, \theta_{c2}, \dots, \theta_{cn})$ is the parameter vector of class c, where n is the number of words in the term document matrix, θ_{ci} is the probability of observing word i in class c and $\sum_i \theta_{ci} = 1$.

Naive Bayes is often used as a baseline in text classification because it is fast and easy to implement. The latter is mainly due to the rather severe assumptions that are made, such as independence of the features. Several problems related to Naïve Bayes classifiers are discussed in Rennie et al. (2003). Here, focus will be on the Complement Naïve Bayes (CNB) classifier, which is more suitable for modelling imbalanced data and features that are not independent. While using only the training data for one specific class, c, when estimating the posterior class probability, P(c|X), in Naïve Bayes classification, the CNB classifier uses information on all classes, except for c. In this way, estimates are more effective because each uses a more even amount of training data per class, which reduces the bias in the weight estimates. The authors obtained more stable weight estimates and improved classification accuracy using this approach. The weights for the decision boundary corresponding to the CNB algorithm correspond to

$$\hat{w}_{\tilde{c}i} = \log \hat{\theta}_{\tilde{c}i}, \text{ with } \hat{\theta}_{\tilde{c}i} = \frac{N_{\tilde{c}i} + \alpha_i}{N_{\tilde{c}} + \alpha},$$

where $N_{\tilde{c}i}$ is the number of times word i occurred in documents in classes other than c and $N_{\tilde{c}}$ is the total number of word occurrences in classes other than c, and α_i and α are smoothing parameters. The classification rule is hence given by

$$l_{CNB}(d) = \arg\max_{c \in C} \left[\log p(\vec{\theta}_c) - \sum_i f_i \log \frac{N_{\tilde{c}i} + \alpha_i}{N_{\tilde{c}} + \alpha} \right].$$

Another way to improve the NB classifier is the Averaged One-Dependence Estimators (AODE) approach, developed by Webb et al. (2005). In AODE, an aggregate of one-dependence classifiers is learned and the final prediction is made by taking the average of the predictions of all these qualified one-dependence classifiers. For simplicity, a one-dependence classifier is firstly built for each attribute, in which the attribute is set to be the parent of all other attributes. Then the AODE directly averages the aggregate consisting of many special tree augmented naïve Bayes. Formally, the classification rule is given by

$$\arg\max_{c \in C} \frac{\sum_{i=1 \wedge F(x_i \geq m)}^{n} P(x_i, c) \prod_{j=1, j \neq i}^{n} P(x_j|x_i, c)}{numParent}$$

where $F(x_i)$ corresponds to the number of training instances having attribute value $x_i$and is used to enforce the lower limit m that is needed to accept a conditional probability estimate, n is the number of attributes, $numParent$ is the number of root attributes which satisfy the condition that the training instances contain at least m examples with the value x_i for the parent attribute X_i. The authors used $m = 30$ in their examples. In addition, Laplace estimates are used for the base probabilities:

$$P(x_i, c) = \frac{F(x_i, c) + 1}{N + v_i * k}$$
$$P(x_j|x_i, c) = \frac{F(x_j, x_i, c) + 1}{F(x_i, c) + v_j}$$

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

where $F()$ is the frequency in the training set of the combination of terms within the brackets, N is the number of training instances, v_i is the number of values of the root attribute X_i, v_j is the number of values of the leaf attribute X_j and k is the number of classes. Jiang and Zhang (2006) provide an extension to the AODE in the sense that different weights are assigned to the distinct tree augmented naïve Bayes, leading to a new algorithm which they called Weightily Averaged One-Dependence Estimators (WAODE). The classification rule is given by

$$\arg\max_{c \in C} \frac{\sum_{i=1}^{n} W_i P(x_i, c) \prod_{j=1, j \neq i}^{n} P(x_j | x_i, c)}{\sum_{i=1}^{n} W_i}$$

where W_i is the weight of the tree augmented naïve Bayes for attribute X_i.

The software that can be used to apply these methods: naiveBayes() function of R package e1017. For the complement naïve Bayes, the AODE and the WAODE, no standard R functions are currently available, but a connection the Weka can be made using the 'RWeka' R package.

2.2.2.1.3 Regression methods

One of the most frequently applied methods for modelling binary outcome variables is the fitting of a logistic regression model, which can be formally stated as

$$P(Y = 1 | X = x) = \frac{\exp(\beta_0 + \beta^t x)}{1 + \exp(\beta_0 + \beta^T x)}$$

$$P(Y = 0 | X = x) = \frac{1}{1 + \exp(\beta_0 + \beta^T x)}$$

Application of the logit transformation (i.e. logit(p) = $\log\{\frac{p}{1-p}\}$), the following linear decision boundary

$$\log \frac{P(Y = 1 | X = x)}{P(Y = 0 | X = x)} = \beta_0 + \beta^T x.$$

The decision boundary is the set of points for which these log-odds are zero:

$$\{x | \beta_0 + \beta^T x = 0\},$$

which can be determined after having obtained estimates for the β coefficients through least squares or maximum likelihood. In addition, shrinkage methods like ridge regression, the Lasso or elastic nets can be employed for restricting the regression model, i.e. to reduce the number of covariates used in the model. In this respect, Dalal et al. (2012) proposed to use Generalized Linear Models with convex penalties (GLMnet), which employs a more general convex penalty that shrinks the coefficients of less important variables to zero, thereby resulting into a model with fewer independent variables that have a better predictive power.

Closely related to these logistic regression models is latent discriminant analysis. In the latter, one assumes a multivariate Gaussian density, with a common covariance matrix, for each of the underlying classes.

Both approaches can be easily extended to include also non-linear decision boundaries.

The software that can be used to apply these methods: glm() function of R package stats, h2o.glm() function of R Package H2O, logistf() function of R package logistf

2.2.2.1.4 K-Nearest-Neighbour and K-means Methods

Nearest-Neighbour methods use those observations in the training set that are closest in input space to the new observation, for which one wants to determine the corresponding class membership. Closeness implies a metric, for which usually the Euclidean distance is considered. Other distance measures can be considered as well (see Hastie et al., 2009).

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

As an example, using the k nearest neighbours approach, we find the k observations with x_i closest to x in input space, and average their responses. Next, in order to determine the class, a majority vote is performed, i.e. if the average is larger than 0.5, the article is considered to be relevant, while the article will be flagged as irrelevant if the average is smaller than 0.5.

Alternatively, k-means clustering aims to partition the observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

The algorithm has a loose relationship to the k-nearest neighbour classifier. One can also apply the 1-nearest neighbour classifier on the cluster centers obtained by k-means to classify new data into the existing clusters. This is known as nearest centroid classifier or Rocchio algorithm, used in 1 of the citations identified by O'Mara-Eves et al. (2015).

The software that can be used to apply these methods: knn() function of R packages class or fnn, kmeans() function of R package stats, kknn() function of R package kknn.

2.2.2.1.5 Classification Trees and Boosting

Classification trees are machine-learning methods for constructing prediction models for categorical data. These models are constructed by recursively partitioning the data space and fitting a simple prediction model within each partition. Hence, they are conceptually simple, yet powerful methods.

In the first step of the process, the algorithm needs to determine which one of the input variables the optimal splitting variable is and, consecutively, which of the values of the selected splitting variable is the optimal split point. Having found the best split, the data are partitioned into the two resulting regions and the splitting process is repeated on each of the two regions. These steps are repeated until the tree is considered to be large enough. Note that the tree size is a tuning parameter that governs the complexity of the model. The optimal tree size can be determined using cost-complexity pruning, a process in which a large tree is pruned until the optimal size. This process is guided by either the misclassification error, the Gini index or the cross-entropy.

In a related fashion, we can also consider the learning idea of boosting. This is a procedure that combines the output of many weak classifiers to produce a powerful committee. A weak classifier ($F(x)$) is one whose error rate is only slightly better than just random guessing and the goal of boosting is to sequentially apply the weak classification algorithm to repeatedly modified versions of the data, thereby producing a sequence of weak classifiers $F_m(x), m = 1,2,\dots,M$. Next, the predictions of all of them are combined through a weighted majority vote to produce the final prediction:

$$F(x) = \text{sign}(\sum_{m=1}^{M} \alpha_m F_m(x)).$$

An example of a weak classifier is a tree with only two terminal nodes, often referred to as a stump. Thinking in terms of the TDM, a stump is based on one specific term in the corpus. If the frequency of that word is smaller than a certain cut-off point, the article is classified as irrelevant, while it is termed relevant if the frequency is higher than the cut-off.

The Gradient Boosting Machine (GBM) applied by Dalal et al. (2012) is a specific implementation of boosting and consists of a general, automated, data-adaptive modelling algorithm that is able to estimate the nonlinear relationship between a variable of interest (in this case whether an article is relevant or not) and a large number of covariates (the frequency of the distinct words in the abstracts, i.e. the TDM) using a sequence of simple classifiers combined in an optimal way.

The software that can be used to apply these methods: rpart() function of R package rpart, ctree() function of R package party, gbm() function of package gbm, h2o.gbm() function of H2O R Package, ada() function of R package ada.

2.2.2.1.6 Neural Networks

A neural network is a two-stage regression or classification model. In case of K-class classification, there are K units at the top of the network, with the kth unit modelling the probability of class k.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Hence there are K target measurements, $Y_k, k = 1, \ldots, K$, each coded as a 0-1 variable for the kth class. Next, derived features Z_m are created from linear combinations of the inputs and the target $Y_k$is modelled as a function of linear combinations of Z_m:

$$\begin{aligned} Z_m &= \sigma(\alpha_{0m} + \alpha_m^T X), m = 1, \ldots, M, \\ T_k &= \beta_{0k} + \beta_k^T Z, k = 1, \ldots, K, \\ f_k(X) &= g_k(T), \quad k = 1, \ldots, K. \end{aligned}$$

The activation function $\sigma(v)$ is usually chosen to be the sigmoid $\sigma(v) = \frac{1}{1+e^{-v}}$. In the current example, we only use 1 hidden layer of units Z_m, but generalisation to a higher number of hidden layers is possible as well. In case the activation function is taken to be the identity function, the entire model collapses to a linear model in the inputs. Therefore, the neural network can be thought of as a non-linear generalization of the standard linear model.

The software that can be used to apply these methods: neuralnet() function of R package neuralnet, h2o.deeplearning()function from H2O R Package.

2.2.2.1.7 Ensemble methods

In the broad sense, ensemble methods refer to approaches that combine two or more machine learning techniques such as the ones discussed above. For example, Dalal et al. (2012) combined the output from their gradient boosting machine and GLMnet approaches into a single classifier, indicating that an article was relevant in case either one of the two classifiers showed a high probability of being relevant. Otherwise stated, an article was considered to be irrelevant only in case both classifiers rejected that article.

Related to the gradient boosting approach described in Section 2.2.2.1.5, is the ensemble method called Random Forests. Instead of using weak classifiers, Random Forest uses fully grown decision trees that are combined later on to make a final decision. Different multiple unpruned decision trees are grown on different bootstrap samples of the training dataset and only a random subset of the features is used to split each node in the decision tree. When it comes down to predicting the final class for an abstract, the class receiving the majority of the votes from the individual trees is considered to be the prediction from the forest.

Another example is the voting perceptron-based automated citation classification system discussed in Cohen et al. (2006). They employed varying learning weights to penalize for misclassification of the relevant class; A drawback is that there is no universal way for choosing the most optimal weights across all reviews beforehand.

Instead of using a combination of different MLTs that can be combined into a single ensemble classifier, one can also use a single MLT and apply it to different feature spaces. This approach was followed in Wallace et al. (2010), who developed an active learning algorithm for SVMs, applied to four different feature spaces. In particular, they used the title text, the abstract text, MeSH keywords (when available) and UMLS terms. For clarity, MeSH (Medical Subject Headings) is a controlled vocabulary thesaurus used for indexing articles for PubMed and the UMLS, or Unified Medical Language System, is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems. In the active learning approach, instead of randomly splitting the data into a training and test set, the system chooses the instances on which it is most confused and requires the human reviewer to label them. As such, only the more informative examples are labelled by humans. This reduces greatly the human workload, but still sustains the required accuracy.

The per-question method described in Frunza et al. (2010) is also an ensemble method that was developed to determine the relevance of specific abstracts. Instead of using the BOW approach, their method tries to provide answers to the questions used by the reviewers to determine relevance of abstracts during the screening process. For each of the 4 questions of interest, a complement naïve Bayes classifier was constructed, and the 4 individual relevance probabilities were then multiplied (so-called voting) to obtain a single indicator of relevance.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Similarly, Kouznetsov and Japkowic (2010) employed a ranking algorithm based on a committee of classifiers. This committee, or ensemble, was constructed using a visualizing classifier performance tool.

Shemilt et al. (2013) proposed another hybrid method that combines the output of the support vector machine classifier with two other methods that require more input from the reviewers. One of these methods requires 'reviewer terms', which are terms that are specified by the review team as being indicative of an includable or excludable study. It is clear that this highly relies on reviewer input. For this reason, their approach will not be further discussed here.

The software that can be used to apply these methods: h2o.randomForest() function from H2O R Package, ranger() function from Ranger R package, Rborist() function from Rborist R package.

2.2.2.1.8 Distributional semantics with relevance feedback

One of the papers identified by O'Mara-Eves et al. (2015) followed a slightly different approach towards determining the relevance of abstracts for a specific systematic review. Instead of restricting to the bag-of-words approach, Jonnalagadda and Petitti (2013) proposed a system that only uses the input and feedback of human reviewers during the course of review. As the reviewers classify articles, the query is modified using a relatively simple relevance feedback algorithm, and the semantically closest document to the query is presented. In this way, the number of articles that needs to be reviewed can be substantially reduced. Nevertheless, since it is not encompassed within the machine-learning environment, this method will not be discussed further in this report.

2.2.2.2 Unsupervised Learning

The supervised learning algorithms discussed in the previous section require at least part of the data to be labelled as either being relevant or irrelevant for the SR to be performed. In contrast, unsupervised learning techniques do not require these labels. Rather, the common aim of these techniques is to group documents that are similar into the same cluster. One of the possible advantages of applying clustering methods is that training data could be selected more adequately. Indeed, instead of taking a random sample of the pool of abstracts, one could consider to perform an initial cluster analysis and select training abstracts from each of the resulting clusters. In this way, it is more likely that a sufficient portion of both relevant and irrelevant papers are used to construct the classifiers. In addition, one can also investigate the different clusters as well to identify what drives these groupings. In an ideal setting, two clusters would be obtained: one corresponding to all relevant papers and one corresponding to all irrelevant papers. Nevertheless, the latter seems to be implausible in real-life SRs.

Kaufman and Rousseeuw (1990) define cluster analysis as the classification of similar objects into groups, where the numbers of groups, as well as their forms are unknown. The "form of a group" refers to the parameters of cluster; that is, to its cluster-specific means, variances, and covariances that also have a geometrical interpretation. Cluster analysis is also called data segmentation. In addition to the grouping or segmenting into subsets or clusters, the goal can be to arrange the clusters into a natural hierarchy, which involves the successively grouping of the clusters themselves so that at each level of the hierarchy, clusters within the same group are more similar to each other than those in different groups. Clustering algorithms fall into three distinct types:

- **combinatorial algorithms**: work directly on the data without underlying probability model,
- **mixture modelling**: supposes data are a sample from some population,
- **mode seeking**: or "bump hunters" take a nonparametric perspective, directly estimating modes of the probability density function.

There are many choices (K-means clustering, K-medoids clustering, hierarchical clustering, agglomerative clustering, divisive clustering, spectral clustering, self-organizing maps, mixtures as soft K-means clustering). More information on these methods is provided in Hastie et al. (2009).

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Analyses can be performed using a variety of R packages, including FactoMineR, cluster, pvclust, mclust, poLCA, bayesclust, bclust, CCMtools, cclust, clue(s), clustersim, fastcluster, ashClust, exclust, modelclust, mclust, pdfCluster, segclust, bpca, etc. Once the clusters are obtained, the reviewer can assess the importance or relevance of the pool of documents more efficiently.

The methods described above, applied in e.g. Stansfield et al. (2013), all make use of the bag-of-words approach, in which the input features are the frequencies of the words in the specific documents. This approach is only effective for grouping related documents in case these documents share a large proportion of lexically equivalent terms. Indeed, instances of synonymy between related documents are ignored. In this respect, Aljaber et al. (2010) present an alternative clustering technique, based on the citation contexts, while Consoli and Stilianakis (2017) use a quartet method based on an input distance matrix derived from XML data returned by PubMed.

2.2.3. Tabulated Overview of the Automation Procedures

The following table presents the identified automation procedures for each of the steps of the SR.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Table 3: Tabulated overview of the Automation Procedures for each of the steps of a systematic review presented in Table 1:

Step/Task	Stage	Automation <u>without</u> MLTs	Automation <u>with</u> MLTs
1. Formulate review question	Preparation	-	-
2. Find previous systematic reviews	Preparation	R packages for retrieving articles from selected databases: • RISmed for PubMed • arXiv for arXiv • rbhl for Biodiversity Heritage Library • rcrossref for Crossref • rdatacite for DataCite • rplos for Public Library of Science Important note: not possible to query all databases and highly dependent on the level of accessibility to the databases of interest	-
3. Write the protocol	Write up	Use of protocol templates (e.g. Cochrane's Review Manager).	-
4. Devise search strategy	Preparation	-	-
5. Search	Retrieval	R packages for retrieving articles from selected databases (same as for step 2) Automatic query expansion algorithms used by most existing databases: synonym expansion and word sense disambiguation. Web-crawling software using Application Program Interfaces (APIs)	-
6. De-duplicate	Retrieval	Reference managers (Mendeley, Endnote, etc.). Endnote + method by Bramer et al. (2016) provides a better performance.	-

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

7. Screen abstracts	Screening	Distributional semantics with relevance feedback. 'Metagear' R package provides tools to initialize a dataframe containing bibliographic data (tile, abstract, journal) from multiple study references, distribute these references randomly to two team members; merge and summarize the screening efforts of this team. It offers a simple tool to quickly run through the abstracts and titles of multiple references and indicate whether the article is relevant. (This is not an automatic procedure.)	Supervised learning, i.e. building classifiers using • Support Vector Machine (SVM) • Naïve Bayes • Regression methods • K-Nearest-Neighbour/K-means • Classification Trees • Boosting • Neural Networks • Ensemble methods (including random forests) Semi-supervised learning: • Label spreading • Label propagation • Semi-supervised support vector machines Unsupervised Learning, I.e. group data into homogeneous clusters. SWIFT-review: workbench with tools for literature prioritisation.
8. Obtain full text	Retrieval	R package 'fulltext' to collect full texts from Biomed Central, Public Library of Science, Pubmed Central, eLife, F1000Research, PeerJ, Pensoft, Hindawi, arXiv and preprints. Of course, open accessibility of the articles is required. R Package 'Metagear' provides also a function to retrieve the full text PDFs (conditional on the journal subscription coverage of the host institution running) Web-crawling software using Application Program Interfaces (APIs)	-
9. Screen full text	Screening	-	There might be a possibility to use the same methods from Step 7, although this has never been described in the existing literature.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

10. Snowball	Retrieval	Modified version of Parscit, using NLP	-
11. Extract data	Synthesis	ExaCt tool: reduce the amount of text to be explored using information highlighting algorithms. 'Metagear' R package: extraction of figures and images from PDFs; web scraping of citations; automated and manual data extraction from scatter-plot and bar-plot images. Shallow semantic parsing, i.e. labelling phrases of a sentence with semantic roles with respect to a target word (natural language processing)	Due to vast amount of different data elements, no universal tool exists. Several attempts in literature for specific applications: • Conditional Random field classifiers (including the BANNER, which was explicitly developed for biomedical text and highly dependent on availability of features known to be related to the topic of interest • Probabilistic graphical models More general approaches exist, making use of • Support Vector Machines • Logistic regression models • Naïve Bayes • Random forests
12. Critical appraisal	Critical Appraisal	-	Identifying certain sources of bias (and extracting sentences to support decision) through the use of • Support Vector Machines • Logistic regression models RobotReviewer: tool developed for assessing bias in randomised trials
13. Synthesize data	Synthesis	Dependent on the research question, data might have to be summarized into a narrative or quantitative synthesis. Data highlighting might prove useful in this respect. For the quantitative synthesis, see step 15 for meta-analysis	-
14. Re-check literature	Retrieval	-	-
15. Meta analyse	Synthesis	Much software for combining and reporting the results from a meta-analysis. General R packages include 'rmeta' and 'metafor', while more specific analyses can be performed with 'epiR', 'exactmeta', Mac and bspmma	
16. Write up review	Write up	-	-

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

2.2.3.1 Discussion of the pros and cons of the distinct learning methods

In order to compare the machine learning techniques that were introduced above, several aspects should be regarded. On the one hand, there are the general, intrinsic properties of the different methods, such as interpretability and complexity. On the other hand, their classification performance, specifically in the field of classification, can be regarded.

2.2.3.1.1 Intrinsic properties of machine learning techniques

From Table 2: above, it became apparent that the SVM was one of the most popular methods to be applied. This approach can cope well with a high number of features (as is the case for abstract screening) and it can also handle irrelevant features. Another advantage is the ability to create both linear and non-linear decision boundaries, making the rule for classification more general. It is also possible to consider linear combinations of features as new input variables for the model. Although the complexity of the SVM is relatively high, it remains easier to interpret compared to the neural networks. One of the major drawbacks, however, is the large amount of memory resources and time that is required to train these models and to classify new observations with it.

A faster computation time is achieved with the classification trees. This approach has also a more intuitive interpretation of the result and it is therefore a lot easier to understand the underlying logic. The danger with this method is to overfit the data, thereby decreasing its classification performance. Also, there is no possibility to consider linear combinations of the inputs and the associated variability can be relatively high. In this respect, the boosting approach can combine several simple trees in more powerful classifier, while also reducing the bias and variance of the individual trees. Other ensemble methods exist and share the advantage that the combined classifier is most often stronger than the individual classifiers. Of course, creating an ensemble is more time consuming and the gain in performance does not always weigh up against the loss in terms of computation time.

Regression methods are also easy and intuitive to interpret and provide smooth and stable classifiers with a low associated variance. Nevertheless, they rely on the assumption of a linear decision bound and can lead to bias.

The problem of bias is less pronounced when using the K-nearest neighbour or means approach. Additional advantages of these approaches include the ease of implementation and interpretation. In addition, there are no underlying assumptions (such as linearity) and they do not rely on specific parameters to be tuned. They are less attractive, however, due to the high computational burden related to calculating the distance between all observations. In addition, accuracy is degraded by the presence of noisy or irrelevant features and might also be influenced greatly by minor changes in the inputs (i.e. become unstable).

In contrast to the nearest neighbour methods, the neural network approach is not sensitive to noisy or contradictory data. In addition, it can deal with high dimensional features and due to the fact that it can fit both linear and non-linear decision boundaries, it is able to construct a powerful classifier. Also, the use of linear combination of inputs can be incorporated. The major reason why they are not used that often for the abstract screening task throughout literature is their high computation cost in combination with the difficulty to understand the underlying logic. The NN is often regarded as a black-box method, that provides nice classification results, but without an intuitive interpretation related neither to its results nor to the way parameters can be dealt with.

The simplicity of the naïve Bayes classifier is in high contrast with the complexity of the neural networks. It is easy to implement and can be constructed relatively fast as it only requires a small amount of training data. Unfortunately, it has a low classification performance and, especially, it relies heavily on the independence assumption, which is almost always violated in abstract screening.

Finally, some attention is paid to two general concepts related to classification through machine learning techniques, i.e. scalability and class imbalance.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Scalability refers to the behaviour MLTs in small versus large datasets. Indeed, a lot of the examples that were found in literature comprise relatively small datasets that are used to construct classifiers and test their performance. Nevertheless, in the context of a systematic review, also large datasets are not uncommon. Especially in the screening stage, the pool of possibly relevant abstracts might be very large. Scalability refers to the effect that an increase in the size of the training set has on the computational performance of the algorithm. In general, it is known that some of the MLTs discussed in their standard forms above cannot handle big datasets well. While the computational performance of some methods is rather robust against changes in sample sizes (e.g. classification trees), others require a modification in the estimation procedure to deal with big datasets (e.g. the stochastic gradient descent procedure for estimating SVMs). The R package 'h2o' () provides parallel distributed machine learning algorithms such as generalized linear models, gradient boosting machines, random forests, and neural networks. This parallel computing greatly improves the speed with which these classifiers are trained.

Class imbalance, on the other hand, refers to unequal proportion of relevant and irrelevant instances in the training set. In most SRs, the number of irrelevant articles is often much higher than the number of relevant ones. Various methods have been proposed to counter this problem of imbalance. Although Frunza et al. (2011) claim that naïve Bayes classifiers perform better on imbalanced datasets compared to classification trees, SVMs and boosting, it is not certain whether this is because of the class imbalance problem or due to other differences between the approaches. Therefore, some remedial measures to deal with the problem of imbalanced datasets could be of interest. For example, different weights can be assigned towards the relevant and irrelevant instances. Also, active learning and under sampling the non-relevant instances could be solutions to the indicated problem. In order to address the problem of class imbalance, two remedial measures will be applied in the case studies. More specifically, SMOTE and ROSE sampling (Lunardon et al., 2014) can be selected to create a more balanced training dataset. In short, SMOTE draws artificial samples by choosing points that lie on the line connecting the rare observation to one of its nearest neighbours in the feature space. ROSE, on the other hand, uses smoothed bootstrapping to draw artificial samples from the feature space neighbourhood around the minority class. Both remedial measures help to relieve the seriousness of the effects of an imbalanced distribution of classes by aiding both the model estimation and model assessment phases and could therefore be recommended in the setting of MLTs in SRs.

Table 4: gives an overview of these basic properties, together with the availability of R packages to perform the analyses.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Table 4: R packages to use for the different machine learning techniques (MLTs) and pros/cons of the different techniques

MLT	R package	Pros	Cons
Support Vector Machines	e1071 kernlab	• Outstanding classification effectiveness • Can handle high-dimensional feature spaces • Cope very well with irrelevant features • Easier to interpret and less prone to overfitting compared to neural networks • Possible to include linear combinations of the input variables • Can fit linear and non-linear decision boundaries	• Relatively complex algorithms • High time and memory consumption during training and classification stage
Naïve Bayes	e1071 RWeka	• Requires small amount of training data • Classification based on probability, hence arrives at correct classification if the correct category is more probable than other categories • Easy to implement	• Low classification performance • Conditional independence assumption often violated in text classification (bad performance of classifier when features are highly correlated)
Regression Methods	stats h2o logistf	• Easy and intuitive interpretation • Smooth and stable classifier • Low associated variance	• Relies heavily on the assumption of a linear decision bound • Might lead to biased results
K-Nearest-Neighbour/K-means	class stats RWeka kknn fnn	• Effective method with low bias • No tuning of parameters • Easy to implement • Easy and intuitive interpretation • No underlying assumptions and can be adapted to specific situations	• Computationally intensive (distance computation for all features), hence runtime performance relatively slow • Accuracy is degraded by presence of noisy or irrelevant features, i.e. high variability • Might become unstable

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Classification trees	rpart party tree h2o gbm ada	• Simple to construct and very intuitive to interpret • Good overview of underlying logic • With 0-1 outcome, computation time is relatively fast	• Often based on very few features, which leads to poor performance in text classification as the number of relevant features is often high • Tendency to over-fit training data • High associated variability, i.e. small changes in data might result into different series of splits • Cannot model additive structure (i.e. combinations of input variables)
Neural Networks	neuralnet h2o	• Can easily handle high-dimensional features • Cope very well with noisy or contradictory data • Very powerful classifier • Can fit linear and non-linear decision boundaries • Possible to use linear combinations of input variables	• High computation cost • Very difficult to understand for average user (black-box method), no intuitive interpretation
Ensemble methods (random forests, boosting)	randomForest adabag h2o	• Combining different classifiers might improve accuracy and performance • Some can reduce computation time or reduce the required amount of training data • Boosting reduces bias and variance • Random forests are very accurate and run efficient on large datasets • Random forests compute proximities between pairs of cases and, as such, can also be used for unsupervised learning	• Not proven to always improve the performance of the individual classifiers. Hence, situation-dependent trade-off between extra computation time and possible improvement • Random forests have been observed to overfit for some datasets with noisy classification/regression tasks • Classifications made by random forests are harder to interpret compared to basic trees • random forests are biased in favour of those attributes with more levels

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

2.2.3.1.2 Classification performance of machine learning techniques

With respect to the classification performance of machine learning techniques, different quantities can be computed and compared. In Table 5: the basic terms true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), are introduced. In terms of abstract screening, a TP refers to a relevant document that was also classified as being relevant, while a true negative is an irrelevant document that was also classified as being irrelevant. On the other hand, the false positives and false negatives correspond to the errors that can be made, where the FP are irrelevant documents being classified as relevant and FN are relevant documents that are falsely classified as irrelevant.

Table 5: 2x2 table of a binary classification problem

Classified category/True category	Relevant	Irrelevant
Relevant	TP	FP
Irrelevant	FN	TN

Most often, specific combinations of the values in Table 5: are used to determine the classifier performance instead of the raw quantities only. Following the results from O'Mara-Eves et al. (2015), the three performance measures that are used throughout the most studies are recall, precision and the F measure. The recall, formally defined as $\frac{TP}{TP+FN}$ resembles the proportion of correctly identified relevant documents positives amongst all truly relevant documents. The precision is the proportion of correctly identified relevant documents amongst all documents that were classified as being relevant, i.e. $\frac{TP}{TP+FP}$. It is clear that the higher the recall and precision, the more adequate a classifier is. Finally, the F-measure combines precision and recall into a single value according to the formula $\frac{(\beta^2+1)TP}{(\beta^2+1)TP+FP+\beta^2FN}$. In other words, it is a weighted harmonic mean where more weight is placed on precision in case $\beta < 1$, while values of $\beta > 1$ indicate recall is more important than precision. Other performance measures that are reported less frequently include amongst other the accuracy and the work saved over sampling (WSS). The former refers to the proportion of agreements to the total number of documents, i.e. $\frac{TP+TN}{TP+FP+TN+FN}$. On the other hand, WSS refers to the percentage of articles that a reviewer does not have to read because they have been screened out by a classifier. At 95% recall, the WSS corresponds to $\frac{FN+TN}{N-0.05}$. Also, the area under the curve (AUC) created by plotting the true positive rate against the false positive rate is sometimes reported as a measure of performance. An AUC equal to 1 corresponds to a perfect score, while 0.5 corresponds to a random ordering. Brodley et al. (2012) argue that is better to use recall and precision over AUC or accuracy due to the fact of having class imbalance (unequal portion of truly relevant and irrelevant articles) and highly asymmetric costs (FN are more detrimental than FP).

According to Sebastiani (2002), different sets of experiments may be used for the comparison of different classifiers, only in case the experiments have been performed:

- on exactly the same collection, i.e. the same documents and the same categories;
- with the same split between training set and test set;
- with the same evaluation measure and, in case this measure depends on parameters, with the same parameter values.

Among the articles that discussed machine learning techniques for the screening of abstracts, none of the above criteria were fulfilled. Only within two of them, the same dataset was used, although no record was made about the choices related to the employed test and training set. Bekhuis and Demner-Fushman (2010) compared the evolutionary support vector machine with the Weightily Averaged One-Dependence Estimators and Naïve Bayes classifier. Based on recall, precision and F1-

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

value, EvoSVM was found to be the superior classifier, followed by WAODE and finally NB. Based on the same performance measures, but a different dataset, SVM and NB were found to perform similar by Frunza *et al*. (2011). From this, it is important to note that applying MLTs to different datasets might lead to different conclusions in terms of classification performance. In addition, the SVM performs good in both studies. The latter is confirmed by the results given in Sebastiani (2002), who focused on the more general problem of classifying documents in multiple categories (i.e. not just screening abstracts for their relevance). The most elaborate study they discussed comprised 12902 documents to be classified into 90 categories. The best performing algorithm was the boosting approach, closely followed by SVM. In a second study, classifying the same documents in only 10 categories identified the SVM as being the most appropriate one. Based upon 5 different comparative studies in this more general classification set-up, Sebastiani (2002) concluded that boosting-based classifier committees, support vector machines and regression methods deliver top-notch performance, followed first by neural networks, and second by the Rocchio and Naïve Bayes classifiers.

Based on these observations, an extensive performance comparison was made based on the three case studies. The focus was addressed towards the support vector machines (with linear, polynomial and radial basis functions), gradient boosting machines, random forests and neural networks. Naïve Bayes had been initially explored as well, but failed to deliver appropriate results. The more in-depth investigation of the performance measures is discussed in Section 3.

3. Results

3.1. Abstract screening

In this section, the results from the three case studies are presented. For each dataset, the same workflow was used. More specifically, the introduced machine learning techniques were trained using training sets of different magnitude. Three specific settings were used, i.e. 20%, 50% and 80% of training data. In each of these settings, the MLTs were trained using no adjustment for imbalanced data, using SMOTE sampling for adjusting the class imbalance and also using the ROSE sampling procedure for correcting the imbalanced data setting. Hence, 18 different (individual) classifiers were constructed.

In addition, as mentioned above in Section 2.2.2, this was done using the 'TDM' input space and the 'Topics' input space. In conclusion, for each dataset, the 18 individual classifiers were trained 6 times. The performance measures for these individual classifiers are presented in Figure 3: - Figure 8: for the Isoflavones case study, Figure 9: - Figure 14: for the QPS case study and Figure 15: - Figure 20: for the ERIS case study. In this perspective, it is important to note that these performance measures are calculated on the test set, i.e. the remaining percentage of the data that was not used for training the classifiers. Of course, since different percentages of the data were used for training, also different percentages of the test set are used. As a result, these individual performance measures are only comparable within a case study and between settings that use the same percentage of training data. In order to solve this issue, different classifiers were selected, saved and tested using a validation set. In total, 4 classifiers were selected based on the following criteria: 1) best F1 value on test set; 2) best sensitivity value on test set; 3) combination of high sensitivity and high specificity (i.e. sometimes sacrificing a small part of the best sensitivity to gain in specificity); and 4) an ensemble of individual classifiers.

The validation set corresponds to 20% of the original data in each case study and this part of the data was not used when constructing any of the MLTs. In this way, the predictions made on this validation set are directly comparable for all classifiers that were constructed within a specific case study. A summary of the selected methods can be found in Table 6: for the Isoflavones case study, in Table 7: for the QPS case study and in Table 8: for the ERIS case study.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

3.1.1. Isoflavones case study

Figure 3: - Figure 8: in Appendix A show the individual performance measures for the 18 classifiers that were introduced above. In this section, a discussion is given on the MLTs that were selected based on the 4 performance criteria. These selected MLTs are presented in Table 6:

A first important conclusion from the observed results is that all the individual classifiers that were selected were based on either the SMOTE or ROSE correction for imbalanced data. This means that ignoring this correction will lead to sub-optimal classification results. In relation to this, it was also noted that classifiers that were trained without correcting for class imbalance can positively contribute to ensembles (e.g. see 20% of training data using the topics feature space).

Secondly, it can be observed that all the classifiers that were selected based on the F1 measure reached a great specificity of at least 90%. On the other hand, the sensitivity measures were rather low, ranging from 63-66% with the TDM feature space and from 51-70% in the topics feature space. As such, this measure is not recommended in the setting of abstract screening. Indeed, the lower the sensitivity, the more relevant papers are incorrectly classified as irrelevant. This is an error that should be minimized.

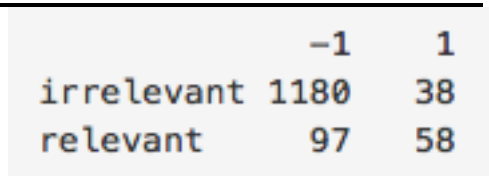
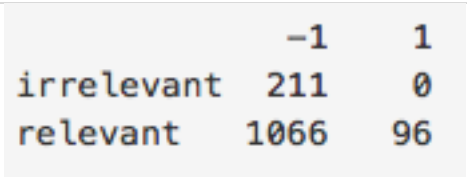
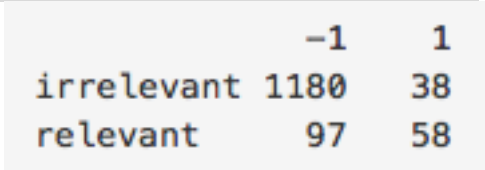
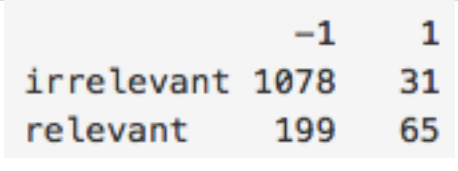
In order to maximize the sensitivity of the classifier, it is also not a good option to just select the classifier with the highest sensitivity on the test set, especially when using the TDM feature space. Indeed, although sensitivity measures of at least 88% could be observed in all settings, the corresponding specificity measures only ranged from 11-35% for the TDM feature space. For the topics feature space, the specificity measures were higher, ranging from 66-68%. The lower the specificity measures, the more irrelevant abstracts are incorrectly classified as being relevant, thereby inducing more work for the reviewers in the phase of full text screening. Rather than using this option, it is more recommended to also account for the specificity measure when selecting the most optimal classifier. Indeed, when sacrificing a little bit of sensitivity, one could increase the specificity and thereby obtain a classifier that has appropriate quantities for both performance measures. Again here, the classifiers based on the topic feature space provide better performance measures, even in the case of only a small training set of 20% of the original data.

The last criterion that was employed is the ensemble of classifiers. It can be observed that even better classifiers can be obtained, also in the case of the TDM feature space. Most promising is the topics feature space, where with using 50% as training data, a sensitivity of 93% could be achieved, while still attaining a specificity of 75%. Adding additional data to the training set only marginally increases the specificity to 77%, while the sensitivity measure remains constant. Of course, the smaller the amounts of training data, the less manual effort for the reviewers as labels for the training data are required. A downside of this ensemble method is the selection of the classifiers it should contain. Not only does the performance of the ensemble depend on the employed methods, but also on the order in which they are included.

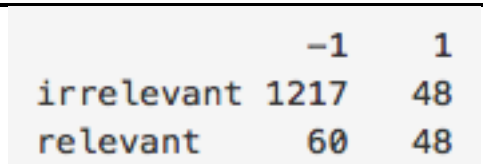
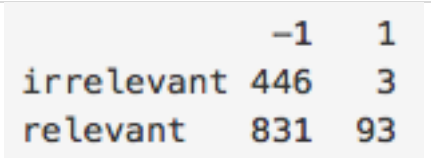
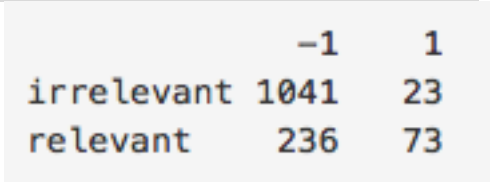
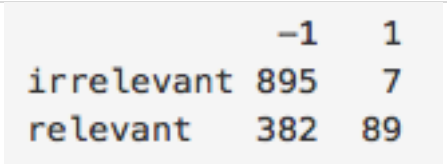
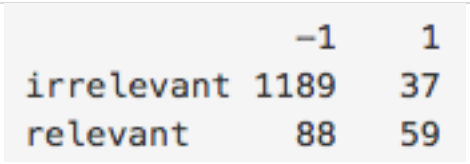
From these results, it is clear that the topics feature approach provides the most promising performance measures, not only on the test set, but also on the validation set. The classifiers that were most commonly selected are the random forests, neural networks and gradient boosting machines. These classifiers are to be used in combination with SMOTE or ROSE sampling to account for class imbalance. Taking into account timing, manual effort and performance, the GBM (with ROSE sampling) training on 50% of the original data using the topics feature space would be selected. Note that this individual classifier is outperformed in terms of specificity by the ensemble that was constructed in the same setting (ensemble of: svm_Linear_smote, svm_Linear_rose, svm_Poly_smote, svm_Poly_rose, svm_Radial_smote, GBM_rose, NN_rose, RF_rose and NN_smote). Nevertheless, this ensemble takes more time to construct and it van be hard to come up with the ideal combination of classifiers. In addition, when regarding the validation set, the performance is the same for the individual GBM compared to the ensemble.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

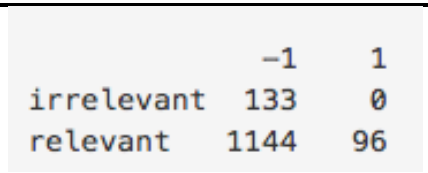
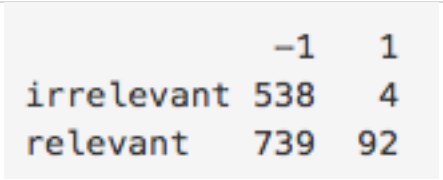
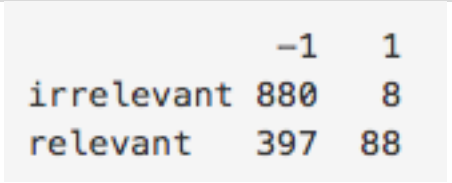
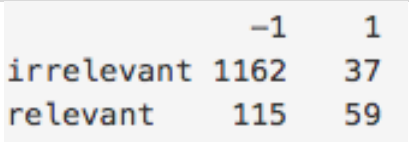
Table 6: Summary of the performance of selected classifiers for the Isoflavones case study

Feature space	Percentage of training	Performance measure	Best MLT	Performance test set	Performance validation set	Predictions validation set[(a)]
TDM	20%	F1 measure	RF_SMOTE	Sensitivity: 0.65 Specificity: 0.92 F1: 0.43	Sensitivity: 0.60 Specificity: 0.92 F1: 0.46	
TDM	20%	Sensitivity	RF_ROSE	Sensitivity: 0.98 Specificity: 0.16 F1: 0.11	Sensitivity: 1 Specificity: 0.17 F1: 0.15	
TDM	20%	Sensitivity + Specificity	RF_SMOTE	Sensitivity: 0.65 Specificity: 0.92 F1: 0.43	Sensitivity: 0.60 Specificity: 0.92 F1: 0.46	
TDM	20%	Ensemble	NN_smote, NN_rose, RF_smote, RF_rose	Sensitivity: 0.74 Specificity: 0.86 F1: 0.35	Sensitivity: 0.68 Specificity: 0.84 F1: 0.36	

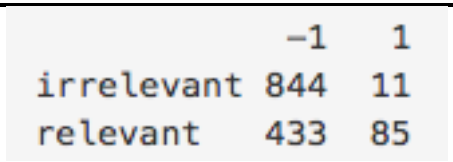
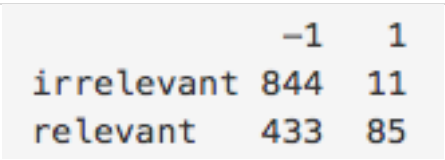
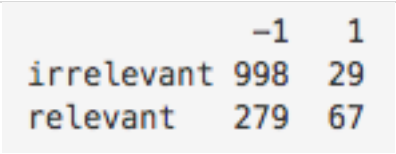
The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Feature space	Percentage of training	Performance measure	Best MLT	Performance test set	Performance validation set	Predictions validation set[(a)]
TDM	50%	F1 measure	RF_SMOTE	Sensitivity: 0.66 Specificity: 0.95 F1: 0.51	Sensitivity: 0.50 Specificity: 0.95 F1: 0.47	
TDM	50%	Sensitivity	RF_ROSE	Sensitivity: 0.99 Specificity: 0.35 F1: 0.14	Sensitivity: 0.97 Specificity: 0.35 F1: 0.18	
TDM	50%	Sensitivity + Specificity	NN_ROSE	Sensitivity: 0.78 Specificity: 0.84 F1: 0.33	Sensitivity: 0.76 Specificity: 0.82 F1: 0.36	
TDM	50%	Ensemble	svm_Linear_smote, GBM_rose, NN_rose, RF_rose	Sensitivity: 0.89 Specificity: 0.71 F1: 0.25	Sensitivity: 0.93 Specificity: 0.70 F1: 0.31	
TDM	80%	F1 measure	GBM_SMOTE	Sensitivity: 0.63 Specificity: 0.95 F1: 0.49	Sensitivity: 0.61 Specificity: 0.93 F1: 0.49	

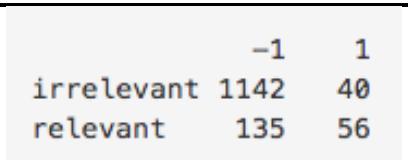
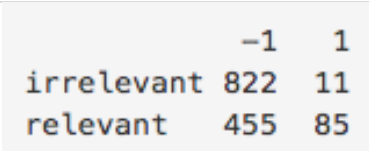
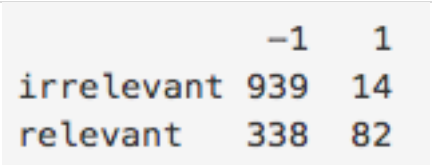
The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Feature space	Percentage of training	Performance measure	Best MLT	Performance test set	Performance validation set	Predictions validation set[(a)]
TDM	80%	Sensitivity	GBM_ROSE	Sensitivity: 1 Specificity: 0.11 F1: 0.10	Sensitivity: 1 Specificity: 0.10 F1: 0.14	
TDM	80%	Sensitivity + Specificity	RF_ROSE	Sensitivity: 0.96 Specificity: 0.44 F1: 0.16	Sensitivity: 0.96 Specificity: 0.42 F1: 0.20	
TDM	80%	Ensemble	RF_rose, svm_Radial_smote, GBM_rose, svm_Linear_rose, NN_rose, svm_Linear_smote	Sensitivity: 0.91 Specificity: 0.69 F1: 0.24	Sensitivity: 0.92 Specificity: 0.69 F1: 0.30	
Topics (30)	20%	F1 measure	RF_SMOTE	Sensitivity: 0.51 Specificity: 0.90 F1: 0.31	Sensitivity: 0.61 Specificity: 0.91 F1: 0.44	

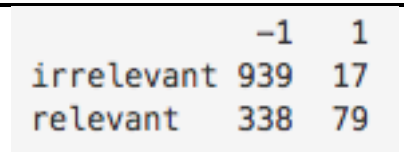
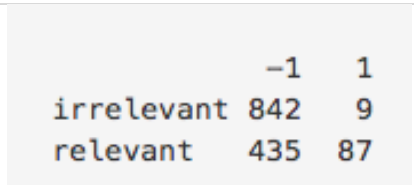
The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Feature space	Percentage of training	Performance measure	Best MLT	Performance test set	Performance validation set	Predictions validation set[(a)]
Topics (30)	20%	Sensitivity	RF_ROSE	Sensitivity: 0.88 Specificity: 0.68 F1: 0.23	Sensitivity: 0.89 Specificity: 0.66 F1: 0.28	
Topics (30)	20%	Sensitivity + Specificity	RF_ROSE	Sensitivity: 0.88 Specificity: 0.68 F1: 0.23	Sensitivity: 0.89 Specificity: 0.66 F1: 0.28	
Topics (30)	20%	Ensemble	svm_Radial_smote, svm_Poly_orig, GBM_orig, svm_Poly_smote, RF_smote, GBM_smote, NN_smote, svm_Linear_smote, NN_rose, svm_Linear_rose, svm_Poly_rose, GBM_rose, RF_rose, svm_Radial_orig, svm_Radial_rose	Sensitivity: 0.76 Specificity: 0.80 F1: 0.28	Sensitivity: 0.70 Specificity: 0.78 F1: 0.30	

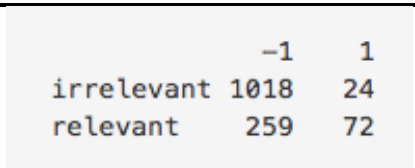
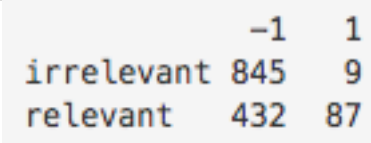
The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Feature space	Percentage of training	Performance measure	Best MLT	Performance test set	Performance validation set	Predictions validation set[(a)]
Topics (30)	50%	F1 measure	GBM_SMOTE	Sensitivity: 0.65 Specificity: 0.90 F1: 0.38	Sensitivity: 0.58 Specificity: 0.89 F1: 0.39	
Topics (30)	50%	Sensitivity	GBM_ROSE	Sensitivity: 0.91 Specificity: 0.66 F1: 0.22	Sensitivity: 0.89 Specificity: 0.64 F1: 0.27	
Topics (30)	50%	Sensitivity + Specificity	NN_ROSE	Sensitivity: 0.87 Specificity: 0.76 F1: 0.28	Sensitivity: 0.85 Specificity: 0.74 F1: 0.32	

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Feature space	Percentage of training	Performance measure	Best MLT	Performance test set	Performance validation set	Predictions validation set[(a)]
Topics (30)	50%	Ensemble	GBM_rose, NN_rose, RF_rose, svm_Poly_rose, svm_Linear_rose, svm_Linear_smote, NN_smote, RF_smote, GBM_smote, svm_Poly_smote, GBM_orig, svm_Poly_orig, svm_Radial_orig, NN_orig	Sensitivity: 0.93 Specificity: 0.75 F1: 0.28	Sensitivity: 0.82 Specificity: 0.74 F1: 0.31	
Topics (30)	80%	F1 measure	RF_SMOTE	Sensitivity: 0.70 Specificity: 0.91 F1: 0.42	Sensitivity: 0.61 Specificity: 0.91 F1: 0.45	
Topics (30)	80%	Sensitivity	GBM_ROSE	Sensitivity: 0.91 Specificity: 0.66 F1: 0.23	Sensitivity: 0.91 Specificity: 0.666 F1: 0.28	

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Feature space	Percentage of training	Performance measure	Best MLT	Performance test set	Performance validation set	Predictions validation set[(a)]
Topics (30)	80%	Sensitivity + Specificity	NN_ROSE	Sensitivity: 0.84 Specificity: 0.79 F1: 0.30	Sensitivity: 0.75 Specificity: 0.80 F1: 0.34	
Topics (30)	80%	Ensemble	svm_Linear_smote, svm_Linear_rose, svm_Poly_smote, svm_Poly_rose, svm_Radial_smote, GBM_rose, NN_rose, RF_rose, NN_smote	Sensitivity: 0.93 Specificity: 0.768 F1: 0.24	Sensitivity: 0.91 Specificity: 0.66 F1: 0.28	

(a): Human decisions: "-1" = excluded, "1" = included. Machine Learning predictions: "irrelevant" or "relevant"

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

3.1.2. QPS case study

Figure 9: - Figure 14: in Appendix A show the individual performance measures for the 18 classifiers that were introduced above. In this section, a discussion is given on the MLTs that were selected based on the 4 performance criteria. These selected MLTs are presented in Table 7:

Similar to the first case study, all individual MLTs that were selected on either one of the first three performance criteria have been trained using additional SMOTE or ROSE sampling to account for class imbalance. The gradient boosting machine was selected only twice, while the most promising MLTs in the current case study are the random forests and neural networks.

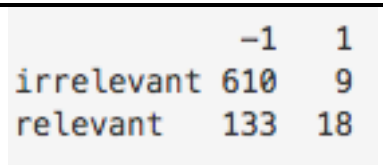
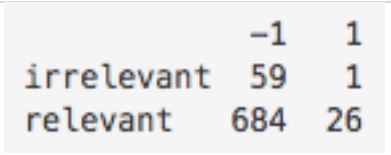
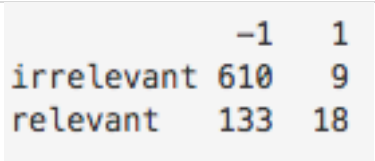
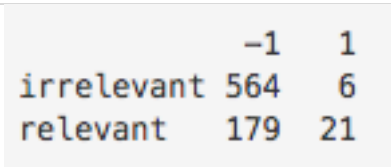
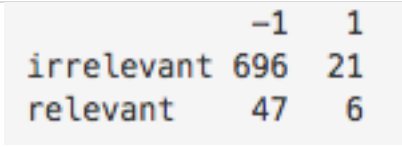
Again here, selecting MLTs based on the F1 measure results into classifiers with a high specificity (all higher than 80%), but with a low to moderate sensitivity (ranging between 37-71% for the TDM feature space and between 44-57% for the topics feature space). Hence, the conclusion from the first case study to not consider this performance measure for the selection of MLTs also holds in the QPS case study. Indeed, for the topics feature space, much better classifiers can be obtained already by regarding the sensitivity on the test set. For the TDM feature space, a combination of the sensitivity and specificity is required to avoid classifiers in which the sensitivity is very high, but the proportion of correctly classified irrelevant articles (i.e. specificity) is too low. Finally, the constructed ensembles in this case study do all provide good performance estimates. Based on only 20% of training data, these ensembles have a sensitivity of 86% for the TDM and 91% for the topics feature space. This means that, on the test set, these percentages of the relevant abstracts are classified as being relevant. On the other hand, the specificity levels are also acceptable, with 75% for the TDM and 52% for the topics feature space. As such, these classifiers are able to reduce the manual work by at least 48%, keeping the error of missing relevant abstracts below 14%. These performance measures are also maintained in the validation set, although a slight drop in performance could be observed there.

Increasing the amount of training data further improves on the performance. When using 80% of training data and the topics feature space, the individual neural network classifier achieved a sensitivity of 88%, with a specificity of 69%. No better individual classifier could be observed with the TDM. Nevertheless, also the TDM feature space provides a good ensemble classifier with 80% of training data with sensitivity and specificity on the test set of 92% and 74%, respectively. Unfortunately, this performance is not reflected in the validation set, where the sensitivity has dropped to 74%. Hence, some overfitting might have occurred in this scenario. In relation to this, it is better to go with the ensemble that was constructed on 20% of training data with the topics feature space. Indeed, although the specificity is only 52%, the performance measures are similar on the test and validation set, thereby providing a better generalisation towards future applications.

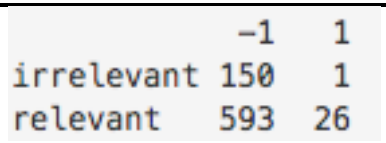
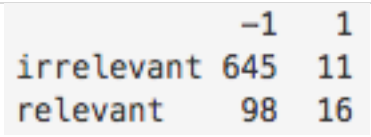
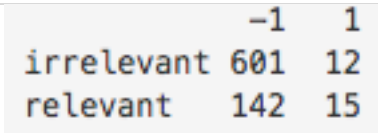
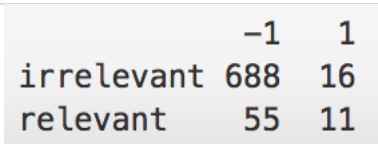
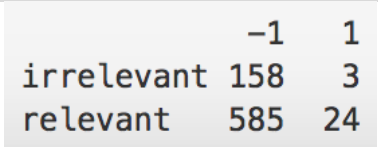
In summary, it can be concluded that also in this second case study, MLTs provide a nice tool to reduce the workload of reviewers in the screening of abstracts phase. Favourable individual classifiers are random forests or neural networks in combination with ROSE or SMOTE sampling, but especially the ensemble of different classifiers seemed to provide the most optimal classifiers.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

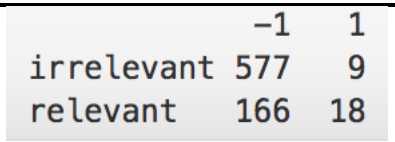
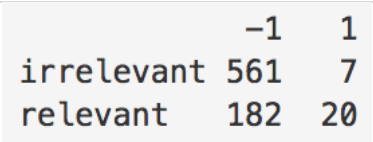
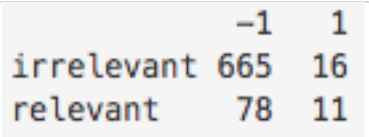
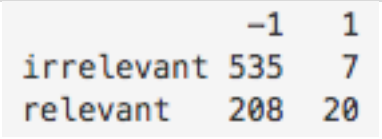
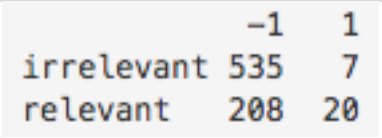
Table 7: Summary of the performance of selected classifiers for the QPS case study

Feature space	Percentage of training	Performance measure	Best MLT	Performance test set	Performance validation set	Predictions validation set[(b)]
TDM	20%	F1 measure	RF_Smote	Sensitivity: 0.71 Specificity: 0.81 F1: 0.22	Sensitivity: 0.67 Specificity: 0.82 F1: 0.2	
TDM	20%	Sensitivity	GBM_Rose	Sensitivity: 0.98 Specificity: 0.09 F1: 0.08	Sensitivity: 0.96 Specificity: 0.08 F1: 0.07	
TDM	20%	Sensitivity + Specificity	RF_Smote	Sensitivity: 0.71 Specificity: 0.81 F1: 0.22	Sensitivity: 0.67 Specificity: 0.82 F1: 0.2	
TDM	20%	Ensemble	RF_smote, svm_Radial_smote, svm_Poly_rose	Sensitivity: 0.86 Specificity: 0.75 F1: 0.22	Sensitivity: 0.78 Specificity: 0.76 F1: 0.19	
TDM	50%	F1 measure	SVM_radial_smote	Sensitivity: 0.37 Specificity: 0.94 F1: 0.26	Sensitivity: 0.22 Specificity: 0.94 F1: 0.15	

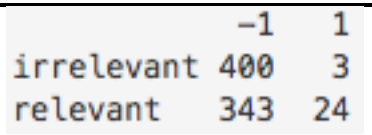
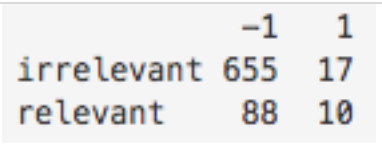
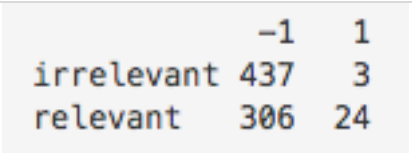
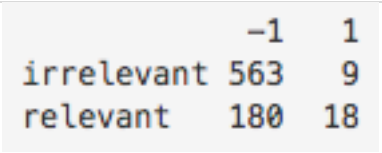
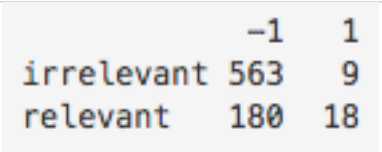
The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Feature space	Percentage of training	Performance measure	Best MLT	Performance test set	Performance validation set	Predictions validation set[(b)]
TDM	50%	Sensitivity	RF_Rose	Sensitivity: 0.95 Specificity: 0.21 F1: 0.09	Sensitivity: 0.96 Specificity: 0.2 F1: 0.08	
TDM	50%	Sensitivity + Specificity	NN_Smote	Sensitivity: 0.63 Specificity: 0.86 F1: 0.25	Sensitivity: 0.59 Specificity: 0.87 F1: 0.23	
TDM	50%	Ensemble	GBM_rose, RF_rose, NN_smote, RF_smote, svm_Linear_rose	Sensitivity: 0.70 Specificity: 0.77 F1: 0.19	Sensitivity: 0.56 Specificity: 0.81 F1: 0.16	
TDM	80%	F1 measure	RF_Smote	Sensitivity: 0.71 Specificity: 0.92 F1: 0.37	Sensitivity: 0.41 Specificity: 0.93 F1: 0.24	
TDM	80%	Sensitivity	RF_Rose	Sensitivity: 0.92 Specificity: 0.26 F1: 0.09	Sensitivity: 0.89 Specificity: 0.21 F1: 0.08	

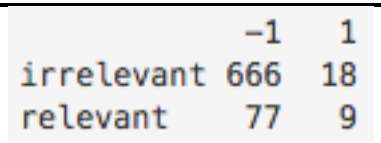
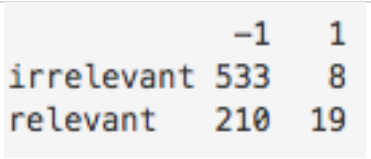
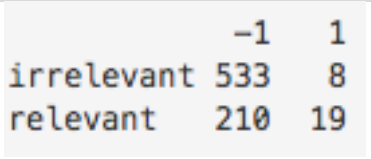
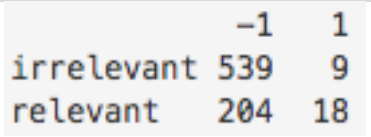
The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Feature space	Percentage of training	Performance measure	Best MLT	Performance test set	Performance validation set	Predictions validation set[(b)]
TDM	80%	Sensitivity + Specificity	NN_Rose	Sensitivity: 0.75 Specificity: 0.80 F1: 0.22	Sensitivity: 0.67 Specificity: 0.78 F1: 0.17	
TDM	80%	Ensemble	NN_rose, RF_smote, NN_smote, svm_Linear_smote, svm_Radial_smote, GBM_orig, GBM_rose, svm_Linear_rose, RF_orig	Sensitivity: 0.92 Specificity: 0.74 F1: 0.22	Sensitivity: 0.74 Specificity: 0.76 F1: 0.17	
Topics (30)	20%	F1 measure	RF_Smote	Sensitivity: 0.44 Specificity: 0.89 F1: 0.21	Sensitivity: 0.41 Specificity: 0.9 F1: 0.19	
Topics (30)	20%	Sensitivity	NN_Rose	Sensitivity: 0.80 Specificity: 0.71 F1: 0.18	Sensitivity: 0.74 Specificity: 0.72 F1: 0.16	
Topics (30)	20%	Sensitivity + Specificity	NN_Rose	Sensitivity: 0.80 Specificity: 0.71 F1: 0.18	Sensitivity: 0.74 Specificity: 0.72 F1: 0.16	

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Feature space	Percentage of training	Performance measure	Best MLT	Performance test set	Performance validation set	Predictions validation set[(b)]
Topics (30)	20%	Ensemble	NN_rose, svm_Linear_orig, GBM_rose, svm_Linear_rose, svm_Radial_orig, svm_Poly_orig	Sensitivity: 0.91 Specificity: 0.52 F1: 0.13	Sensitivity: 0.89 Specificity: 0.54 F1: 0.12	
Topics (30)	50%	F1 measure	GBM_Smote	Sensitivity: 0.57 Specificity: 0.90 F1: 0.29	Sensitivity: 0.37 Specificity: 0.88 F1: 0.16	
Topics (30)	50%	Sensitivity	RF_Rose	Sensitivity: 0.87 Specificity: 0.58 F1: 0.14	Sensitivity: 0.89 Specificity: 0.59 F1: 0.13	
Topics (30)	50%	Sensitivity + Specificity	SVM_Linear_Rose	Sensitivity: 0.70 Specificity: 0.75 F1: 0.18	Sensitivity: 0.67 Specificity: 0.76 F1: 0.16	
Topics (30)	50%	Ensemble	SVM_Linear_Rose	Sensitivity: 0.70 Specificity: 0.75 F1: 0.18	Sensitivity: 0.67 Specificity: 0.76 F1: 0.16	

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Feature space	Percentage of training	Performance measure	Best MLT	Performance test set	Performance validation set	Predictions validation set[(b)]
Topics (30)	80%	F1 measure	RF_Smote	Sensitivity: 0.54 Specificity: 0.90 F1: 0.28	Sensitivity: 0.33 Specificity: 0.9 F1: 0.16	
Topics (30)	80%	Sensitivity	NN_Rose	Sensitivity: 0.88 Specificity: 0.69 F1: 0.19	Sensitivity: 0.7 Specificity: 0.72 F1: 0.15	
Topics (30)	80%	Sensitivity + Specificity	NN_Rose	Sensitivity: 0.88 Specificity: 0.69 F1: 0.19	Sensitivity: 0.7 Specificity: 0.72 F1: 0.15	
Topics (30)	80%	Ensemble	svm_Linear_rose, svm_Poly_orig, GBM_smote	Sensitivity: 0.83 Specificity: 0.74 F1: 0.20	Sensitivity: 0.67 Specificity: 0.73 F1: 0.14	

(b): Human decisions: "-1" = excluded, "1" = included. Machine Learning predictions: "irrelevant" or "relevant"

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

3.1.3. ERIS case study

Individual performance measures for the 18 constructed classifiers are shown in Appendix A (Figure 15: - Figure 20:). In this section, a discussion is given on the MLTs that were selected based on the 4 performance criteria. These selected MLTs are presented in Table 8:

Among the three case studies that were considered, the ERIS case study contained the smallest number of abstracts (668), but was also slightly more balanced (16.77% relevant abstracts). Nevertheless, it could also be observed from Table 8: that, again here, the SMOTE and ROSE sampling was required to provide the most optimal classifiers.

Regarding the F1 measure for selecting classifiers, the conclusion from the former two case studies only holds for the TDM feature space. Indeed, for all percentages of the training data, the selected MLTs had high specificities ranging from 85-94%, but very low sensitivities between 33 and 35%. For the topics feature space, the selected MLTs had better sensitivities which were close to optimal in the current case study. Nevertheless, these sensitivities were lower as compared to the first two case studies as they ranged between 54 and 73%. This is probably the effect of the reduced amount of data.

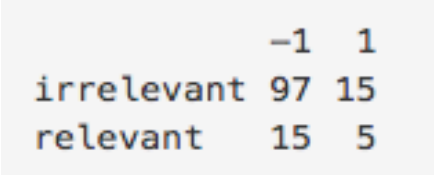
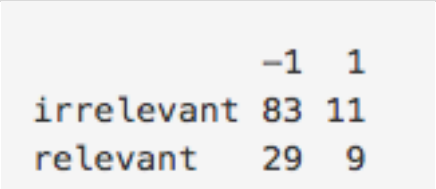
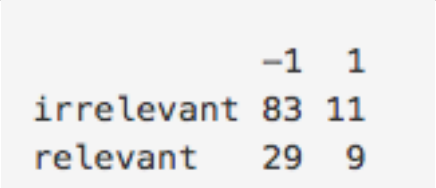
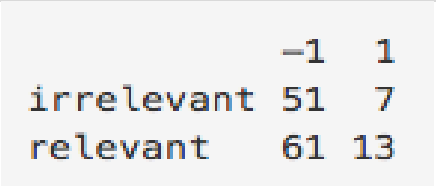
For the topics feature space, the performance measures for the selected classifiers based on any of the criteria were very consistent and appropriate. The TDM feature space, on the other hand, provided less adequate estimates in this case study, with sensitivity measures below 60%. Nevertheless, even for the TDM feature space, considering ensembles of individual classifiers provides adequate performance measures for both the sensitivity and specificity, given that at least 80% of training data is used. This makes sense because the overall amount of available data is rather limited and enough information is required to train the MLTs. Using this amount of training data, the ensemble with the TDM feature space achieves a sensitivity of 83%, with a specificity of 54%, while the topics feature space achieves 89% sensitivity with the same specificity. Hence, these classifiers are able to reduce the manual workload with 46%, while keeping the error rate of missing relevant abstracts below 17%.

Compared to the former two case studies, all individual classifiers have been selected based on one of the selection criteria. Indeed, next to the random forest and neural networks (which were also selected in the Isoflavones and QPS case study), also the support vector machines were occasionally selected. Especially the SVM with a linear kernel in combination with the topics feature space showed promising performance measures.

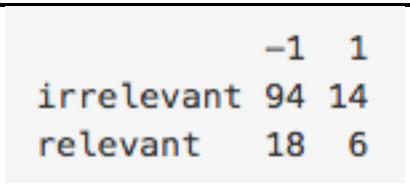
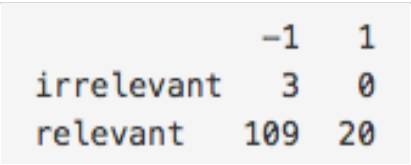
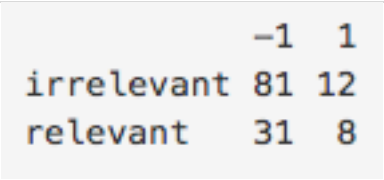
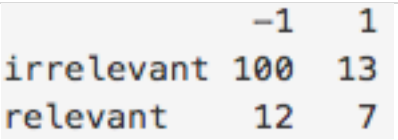
While the conclusions from this case study are somewhat less pronounced compared to the former two, it can be stated that again the random forests in combination with the topics feature space are good individual MLTs in terms of achieving appropriate sensitivity and specificity levels and that especially the ensemble methods (for both the TDM and topics feature spaces) provide a good way of aiding reviewers in the screening of abstracts.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

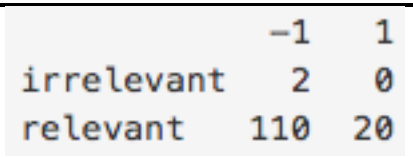
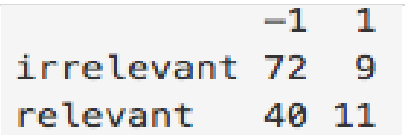
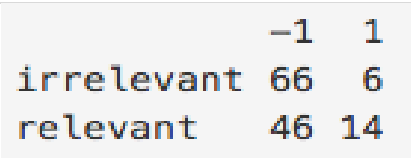
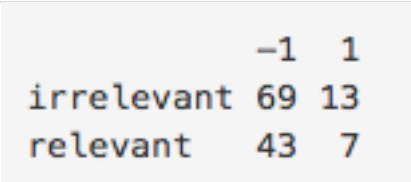
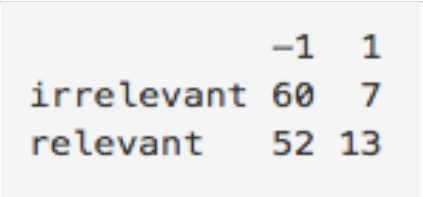
Table 8: Summary of the performance of selected classifiers for the ERIS case study

Feature space	Percentage of training	Performance measure	Best MLT	Performance test set	Performance validation set	Predictions validation set[(c)]
TDM	20%	F1 measure	SVM_Poly_SMOTE	Sensitivity: 0.35 Specificity: 0.85 F1: 0.34	Sensitivity : 0.25 Specificity: 0.87 F1: 0.25	
TDM	20%	Sensitivity	NN_SMOTE	Sensitivity: 0.40 Specificity: 0.76 F1: 0.32	Sensitivity: 0.45 Specificity: 0.74 F1: 0.31	
TDM	20%	Sensitivity + Specificity	NN_SMOTE	Sensitivity: 0.40 Specificity: 0.76 F1: 0.32	Sensitivity: 0.45 Specificity: 0.74 F1: 0.31	
TDM	20%	Ensemble	NN_smote, svm_Poly_smote, GBM_orig	Sensitivity: 0.68 Specificity: 0.46 F1: 0.32	Sensitivity: 0.65 Specificity: 0.46 F1: 0.28	

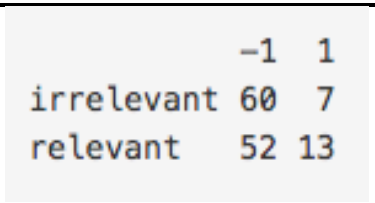
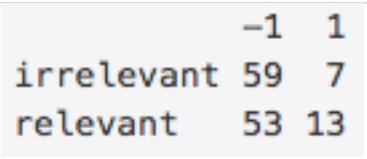
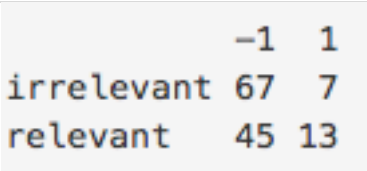
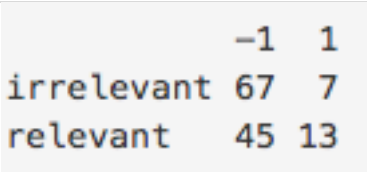
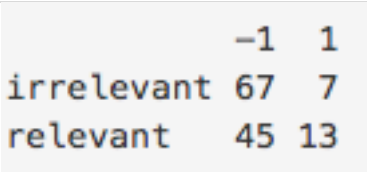
The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Feature space	Percentage of training	Performance measure	Best MLT	Performance test set	Performance validation set	Predictions validation set[(c)]
TDM	50%	F1 measure	GBM_SMOTE	Sensitivity: 0.33 Specificity: 0.91 F1: 0.38	Sensitivity: 0.3 Specificity: 0.84 F1: 0.27	
TDM	50%	Sensitivity	GBM_ROSE	Sensitivity: 0.96 Specificity: 0.06 F1: 0.29	Sensitivity: 1 Specificity: 0.03 F1: 0.27	
TDM	50%	Sensitivity + Specificity	NN_ROSE	Sensitivity: 0.40 Specificity: 0.79 F1: 0.33	Sensitivity: 0.4 Specificity: 0.72 F1: 0.27	
TDM	50%	Ensemble	NN_smote, svm_Linear_rose, GBM_smote	Sensitivity: 0.64 Specificity: 0.65 F1: 0.39	Sensitivity: 0.5 Specificity: 0.64 F1: 0.29	
TDM	80%	F1 measure	GBM_SMOTE	Sensitivity: 0.33 Specificity: 0.94 F1: 0.41	Sensitivity: 0.35 Specificity: 0.89 F1: 0.36	

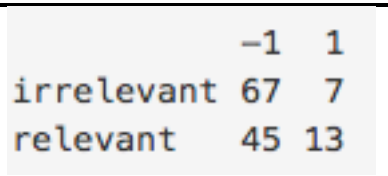
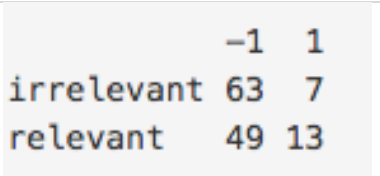
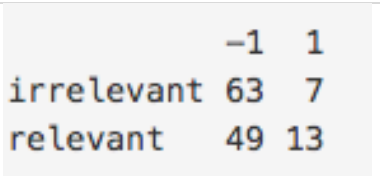
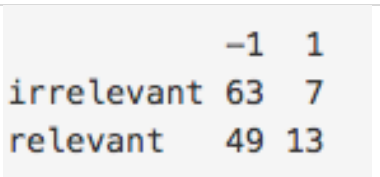
The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Feature space	Percentage of training	Performance measure	Best MLT	Performance test set	Performance validation set	Predictions validation set[(c)]
TDM	80%	Sensitivity	GBM_ROSE	Sensitivity: 1 Specificity: 0.01 F1: 0.30	Sensitivity: 1 Specificity: 0.02 F1: 0.27	
TDM	80%	Sensitivity + Specificity	SVM_Linear_ROSE	Sensitivity: 0.56 Specificity: 0.68 F1: 0.36	Sensitivity: 0.55 Specificity: 0.64 F1: 0.31	
TDM	80%	Ensemble	NN_smote, svm_Linear_rose, RF_smote, GBM_smote, RF_rose	Sensitivity: 0.83 Specificity: 0.54 F1: 0.41	Sensitivity: 0.7 Specificity: 0.59 F1: 0.35	
Topics (30)	20%	F1 measure	SVM_Poly_Rose	Sensitivity: 0.54 Specificity: 0.67 F1: 0.35	Sensitivity: 0.35 Specificity: 0.62 F1: 0.2	
Topics (30)	20%	Sensitivity	RF_ROSE	Sensitivity: 0.64 Specificity: 0.50 F1: 0.32	Sensitivity: 0.65 Specificity: 0.54 F1: 0.31	

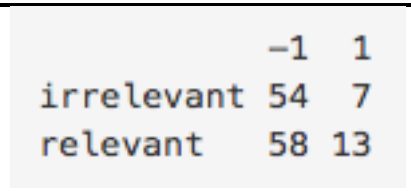
The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Feature space	Percentage of training	Performance measure	Best MLT	Performance test set	Performance validation set	Predictions validation set[(c)]
Topics (30)	20%	Sensitivity + Specificity	RF_ROSE	Sensitivity: 0.64 Specificity: 0.50 F1: 0.32	Sensitivity: 0.65 Specificity: 0.54 F1: 0.31	
Topics (30)	20%	Ensemble	svm_Poly_rose, GBM_rose, RF_rose	Sensitivity: 0.64 Specificity: 0.58 F1: 0.35	Sensitivity: 0.65 Specificity: 0.53 F1: 0.3	
Topics (30)	50%	F1 measure	SVM_Linear_SMOTE	Sensitivity: 0.73 Specificity: 0.69 F1: 0.45	Sensitivity: 0.65 Specificity: 0.6 F1: 0.33	
Topics (30)	50%	Sensitivity	SVM_Linear_SMOTE	Sensitivity: 0.73 Specificity: 0.69 F1: 0.45	Sensitivity: 0.65 Specificity: 0.6 F1: 0.33	
Topics (30)	50%	Sensitivity + Specificity	SVM_Linear_SMOTE	Sensitivity: 0.73 Specificity: 0.69 F1: 0.45	Sensitivity: 0.65 Specificity: 0.6 F1: 0.33	

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Feature space	Percentage of training	Performance measure	Best MLT	Performance test set	Performance validation set	Predictions validation set[(c)]
Topics (30)	50%	Ensemble	SVM_Linear_SMOTE	Sensitivity: 0.73 Specificity: 0.69 F1: 0.45	Sensitivity: 0.65 Specificity: 0.6 F1: 0.33	
Topics (30)	80%	F1 measure	RF_ROSE	Sensitivity: 0.72 Specificity: 0.63 F1: 0.41	Sensitivity: 0.65 Specificity: 0.56 F1: 0.32	
Topics (30)	80%	Sensitivity	RF_ROSE	Sensitivity: 0.72 Specificity: 0.63 F1: 0.41	Sensitivity: 0.65 Specificity: 0.56 F1: 0.32	
Topics (30)	80%	Sensitivity + Specificity	RF_ROSE	Sensitivity: 0.72 Specificity: 0.63 F1: 0.41	Sensitivity: 0.65 Specificity: 0.56 F1: 0.32	

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Feature space	Percentage of training	Performance measure	Best MLT	Performance test set	Performance validation set	Predictions validation set[(c)]
Topics (30)	80%	Ensemble	svm_Radial_smote, svm_Linear_rose, svm_Poly_rose, GBM_rose, NN_smote, NN_rose, svm_Linear_smote	Sensitivity: 0.89 Specificity: 0.54 F1: 0.43	Sensitivity: 0.65 Specificity: 0.48 F1: 0.29	

(c): Human decisions: "-1" = excluded, "1" = included. Machine Learning predictions: "irrelevant" or "relevant"

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

3.1.4. Intermediate conclusion and aspects of computation time

The previous three subsections compared the introduced classifiers in terms of the performance measure F1, sensitivity and specificity when using the different case studies. Table 9 presents a summary of the top best models that were selected in each of the case studies. Up to 4 models are presented, depending on how many would be really recommended.

Table 9: Summary of the recommended models in each of the case studies

	Recommended models			
Case study (amount of data in training set, i.e. full data without validation set)	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Isoflavones (N=5494)	Topics + Gradient boosting machine + ROSE	Topics + Random forest + SMOTE	Topics + Neural network + SMOTE	Topics or TDM in combination with an ensemble of several classifiers
QPS (N=3078)	Topics + Neural network +ROSE	TDM + Random forest + SMOTE	Topics or TDM in combination with an ensemble of several classifiers	–
ERIS (N=527)	Topics + Random forest + ROSE	Topics or TDM in combination with an ensemble of several classifiers	–	–

For the Isoflavones and QPS case study, especially the topics approach could be recommended in combination with random forests, neural networks or gradient boosting machines (each adjusted for class imbalance via SMOTE or ROSE sampling). For the ERIS case study, the observed performance measures were lower and the difference in performance between the TDM or topics feature space was also smaller. Nevertheless, it could be advised as well to consider random forests in combination with the topics feature space or an ensemble trained on either the TDM or topics feature space.

In addition to these performance measures, it is also of interest to regard the computation time associated with each combination of 1) amount of training data and 2) the selected feature space. The table below gives an overview of the computation time related to the construction of all 18 classifiers for each of the considered case studies. The default setting of the tuning parameters in the shiny application was used. Decreasing the number of values in the grid for the tuning parameters will of course decrease computation time.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Table 10: Computation time related to the construction of the 18 classifiers in each of the case studies

		Amount of training data used		
Case study (amount of data in training set, i.e. full data without validation set)	**Feature space (pre-processing time)**	**20%**	**50%**	**80%**
Isoflavones	TDM (few seconds)	21.44 min	1.55 hours	4.4 hours
(N=5494)	Topics (15 min)	4 min	12 min	41 min
QPS	TDM (few seconds)	8.67 min	29.45 min	1.4 hours
(N=3078)	Topics (5 min)	1.8 min	5.4 min	10 min
ERIS	TDM (few seconds)	2.46 min	5.89 min	9.77 min
(N=527)	Topics (1.5 min)	0.73 min	1.36 min	2.12 min

A first observation is related to the pre-processing time, which is longer for the topics approach. Pre-processing refers to all operations that are performed on the raw texts to transform them into an appropriate input space. In addition to the standard pre-processing operations such as stemming and removing stopwords (see Section 2.2.2), applied for the TDM approach, the topics approach first needs to identify topics as well. This can take up to 15 minutes for the largest case study that was explored here (5494 abstracts). Nevertheless, once these topics are identified, the training of the classifiers goes much faster. Therefore, the overall time spent on training all classifiers is much smaller compared to those based on the TDM approach. Together with the performance measures in the previous subsections, this adds to the recommendation of the topics feature space.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

3.2. Initial explorations of other options

3.2.1. Full text screening

In the previous subsection, the focus was on the application of the MLTs in the setting of abstract screening. The introduced methods can however also be applied to a corpus based on the entire article, thereby performing full text screening. Due to limited availability of full texts for all case studies, a small-scale exercise was done based on 90 selected articles from the ERIS case study. For these 90 articles, the pdf files of the entire articles were searched and downloaded from several databases. As discussed during the interim meetings, a balanced setting was considered, with 40 relevant articles and 50 irrelevant articles.

A comparison was made between training on these 90 full texts versus training on the 90 corresponding abstracts. In each setting, 70% of training data was used and the 30% that remained was used as test set. In order to make an adequate comparison, a validation set was considered, consisting of 569 abstracts of remaining articles in the ERIS case study. The results are presented below.

3.2.1.1 Training based on abstracts only

Based on the results in Section 3.1, only the topics feature space was regarded for the current exercise. The performance of the individual MLTs is summarized in Figure 1:

Overview of performance measures for the individual classifiers

Show 25 entries Search:

	Accuracy	Kappa	F1	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
svm_Linear_orig	0.62962962962963	0.23728813559322	0.545454545454545	0.5	0.733333333333333	0.6	0.647058823529412	0.6	0.5
svm_Linear_smote	0.666666666666667	0.307692307692308	0.571428571428572	0.5	0.8	0.666666666666667	0.666666666666667	0.666666666666667	0.5
svm_Linear_rose	0.740740740740741	0.479338842975207	0.72	0.75	0.733333333333333	0.692307692307692	0.785714285714286	0.692307692307692	0.75
svm_Poly_orig	0.481481481481481	-0.0677966101694916	0.363636363636364	0.333333333333333	0.6	0.4	0.529411764705882	0.4	0.333333333333333
svm_Poly_smote	0.592592592592593	0.168067226890756	0.521739130434783	0.5	0.666666666666667	0.545454545454545	0.625	0.545454545454545	0.5
svm_Poly_rose	0.555555555555556	0.1	0.5	0.5	0.6	0.5	0.6	0.5	0.5
svm_Radial_orig	0.555555555555556	0		0	1		0.555555555555556		0
svm_Radial_smote	0.444444444444444	0	0.615384615384615	1	0	0.444444444444444		0.444444444444444	1
svm_Radial_rose	0.555555555555556	0		0	1		0.555555555555556		0
GBM_orig	0.555555555555556	0.0689655172413792	0.4	0.333333333333333	0.733333333333333	0.5	0.578947368421053	0.5	0.333333333333333
GBM_smote	0.592592592592593	0.168067226890756	0.521739130434783	0.5	0.666666666666667	0.545454545454545	0.625	0.545454545454545	0.5
GBM_rose	0.518518518518518	0.120300751879699	0.648648648648649	1	0.133333333333333	0.48	1	0.48	1
NN_orig	0.666666666666667	0.27027027027027	0.4	0.25	1	1	0.625	1	0.25
NN_smote	0.777777777777778	0.526315789473684	0.666666666666667	0.5	1	1	0.714285714285714	1	0.5
NN_rose	0.592592592592593	0.195121951219512	0.592592592592593	0.666666666666667	0.533333333333333	0.533333333333333	0.666666666666667	0.533333333333333	0.666666666666667
RF_orig	0.666666666666667	0.307692307692308	0.571428571428572	0.5	0.8	0.666666666666667	0.666666666666667	0.666666666666667	0.5
RF_smote	0.703703703703704	0.379310344827586	0.6	0.5	0.866666666666667	0.75	0.684210526315789	0.75	0.5
RF_rose	0.62962962962963	0.196428571428571	0.375	0.25	0.933333333333333	0.75	0.608695652173913	0.75	0.25

Showing 1 to 18 of 18 entries Previous 1 Next

Figure 1: Performance measures of the individual classifiers trained on the abstracts of 70% of the selected 90 articles in the ERIS case study, with topics as feature space

The support vector machine with linear kernel and a ROSE correction for imbalancedness was selected as the optimal MLT based on both the F1 measure and the combination of sensitivity and specificity. Note that the increased F1 measures, compared to the results in Section 3.1 are due to the more balanced scenario that was considered here. On the test set, this MLT had a sensitivity of 75% and a specificity of 73%. On the validation set, however, a sensitivity measure of only 56% was obtained, with a specificity

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

of 65%. Similarly, the best performing ensemble was constructed, containing SVM_Linear_Rose, NN_Rose, RF_Rose, NN_Original, SVM_Linear_Orig, NN_Smote and RF_Orig. On the test set, the sensitivity had increased to 92%, with a specificity of 53%. Using this ensemble on the validation set, the sensitivity was almost perfect as only 1 of the 70 relevant abstracts was incorrectly classified as irrelevant. Unfortunately, the specificity was only 15%, meaning that most of the irrelevant abstracts were classified as being relevant as well.

3.2.1.2 Training based on full texts

Again here, only the topics feature space was regarded for the current exercise. The performance of the individual MLTs is summarized in Figure 2:

Overview of performance measures for the individual classifiers

Show 25 entries Search:

	Accuracy	Kappa	F1	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Prec
svm_Linear_orig	0.518518518518518	-0.0733944954128442		0	0.933333333333333	0	0.538461538461539	
svm_Linear_smote	0.740740740740741	0.461538461538462	0.666666666666667	0.583333333333333	0.866666666666667	0.777777777777778	0.722222222222222	0.777777777
svm_Linear_rose	0.518518518518518	-0.0733944954128442		0	0.933333333333333	0	0.538461538461539	
svm_Poly_orig	0.592592592592593	0.108108108108108	0.266666666666667	0.166666666666667	0.933333333333333	0.666666666666667	0.583333333333333	0.666666666
svm_Poly_smote	0.666666666666667	0.307692307692308	0.571428571428572	0.5	0.8	0.666666666666667	0.666666666666667	0.666666666
svm_Poly_rose	0.666666666666667	0.295652173913043	0.526315789473684	0.416666666666667	0.866666666666667	0.714285714285714	0.65	0.714285714
svm_Radial_orig	0.518518518518518	-0.0733944954128442		0	0.933333333333333	0	0.538461538461539	
svm_Radial_smote	0.444444444444444	0	0.615384615384615	1	0	0.444444444444444		0.444444444
svm_Radial_rose	0.518518518518518	-0.0733944954128442		0	0.933333333333333	0	0.538461538461539	
GBM_orig	0.62962962962963	0.210526315789474	0.444444444444444	0.333333333333333	0.866666666666667	0.666666666666667	0.619047619047619	0.666666666
GBM_smote	0.62962962962963	0.210526315789474	0.444444444444444	0.333333333333333	0.866666666666667	0.666666666666667	0.619047619047619	0.666666666
GBM_rose	0.407407407407407	-0.0746268656716417	0.578947368421053	0.916666666666667	0	0.423076923076923	0	0.423076923
NN_orig	0.555555555555556	0		0	1		0.555555555555556	
NN_smote	0.592592592592593	0.123893805309734	0.352941176470588	0.25	0.866666666666667	0.6	0.590909090909091	
NN_rose	0.777777777777778	0.564516129032258	0.785714285714286	0.916666666666667	0.666666666666667	0.6875	0.909090909090909	
RF_orig	0.592592592592593	0.123893805309734	0.352941176470588	0.25	0.866666666666667	0.6	0.590909090909091	
RF_smote	0.62962962962963	0.23728813559322	0.545454545454545	0.5	0.733333333333333	0.6	0.647058823529412	
RF_rose	0.740740740740741	0.48780487804878	0.740740740740741	0.833333333333333	0.666666666666667	0.666666666666667	0.833333333333333	0.666666666

Showing 1 to 18 of 18 entries Previous 1 Next

Figure 2: Performance measures of the individual classifiers trained on the full texts of 70% of the selected 90 articles in the ERIS case study, with topics as feature space

The neural network with a ROSE correction for imbalance was found the be the most optimal classifier and also the most optimal ensemble. The performance measures on the test set were a sensitivity of 92% and a specificity of 67%. Note that this test set is also composed of the full texts of the remaining 30% of the selected 90 articles. On the other hand, the validation set is composed of only the abstracts of the articles. On this validation set, a sensitivity measure of 87% was obtained, with a specificity of 30%. In this perspective, the amount of correctly identified relevant abstracts is somewhat smaller than in the abstract screening part, but the amount of correctly identified irrelevant articles is higher as well. As such, there is no clear winner in this exercise as to which method is preferred (full text or abstracts screening). This is also due to the fact of the limited amount of available training data and a more in-depth study is recommended for the future.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

3.2.2. Predictions on new data

The ultimate goal of the developed application is to replace one of the reviewers in a future systematic review. Indeed, based on historical data, the classifiers can be trained and used in future reviews to be performed on similar data. With this aim, EFA provided new data related to the ERIS case study, consisting of 319 abstracts (285 irrelevant, 34 relevant). These new abstracts are only related to salmon and hence constitute only part of the training data, which contained articles on both salmon and oyster. For illustrative purposed, the best performing classifiers from Section 3.1.3 were used to predict the relevance of these new abstracts. The results are summarised below in Table 11: . The 'classifier' column indicated the employed classifier and feature space, while the 'Performance' column shows the confusion matrix where the predictions are shown in the rows (irrelevant, relevant) and the results from a reviewer at EFSA are shown in the columns (N = irrelevant, Y=relevant).

Table 11: Summary of the performance of selected classifiers, applied to new ERIS data

Classifier	Performance[(d)]
Feature space: TDM Amount of training data: 80% Model: ensemble of NN_smote, svm_Linear_rose, RF_smote, GBM_smote and RF_rose	N Y irrelevant 168 11 relevant 116 23
Feature space: Topics Amount of training data: 80% Model: ensemble of svm_Radial_smote, svm_Linear_rose, svm_Poly_rose, GBM_rose, NN_smote, NN_rose and svm_Linear_smote	N Y irrelevant 124 10 relevant 160 24
Feature space: Topics Amount of training data: 80% Model: RF with rose sampling	N Y irrelevant 185 12 relevant 99 22

(d): Human decisions: "N" = excluded, "Y" = included. Machine Learning predictions: "irrelevant" or "relevant"

It can be observed that all three models are able to identify at least 65% of the relevant abstracts. The workload is most reduced with the random forest approach using ROSE sampling on the topics feature space, closely followed by the ensemble trained on the TDM feature space.

If we retrain the models on a higher percentage of the original data (more precisely, 90% of all original ERIS data), even better classifiers could be constructed. Indeed an ensemble of svm_Linear_rose, GBM_smote and NN_rose with 90% sensitivity and 62% specificity was found using the TDM feature space (computation time 14 minutes). Using the topics feature space (30 topics), an ensemble with RF_rose and RF_smote was constructed that had 82% sensitivity and 71% specificity (computation time 5 minutes). Using these updated models to predict the relevance of the new abstracts resulted into the following confusion matrices.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Table 12: Summary of the performance of updated classifiers, applied to new ERIS data

Classifier	Performance[(e)]
Feature space: TDM Amount of training data: 90% of all ERIS data Model: ensemble of svm_Linear_rose, GBM_smote and NN_rose	N Y irrelevant 140 13 relevant 144 21
Feature space: Topics Amount of training data: 90% of all ERIS data Model: ensemble of svm_Radial_smote, svm_Linear_rose, svm_Poly_rose, GBM_rose, NN_smote, NN_rose and svm_Linear_smote	N Y irrelevant 172 18 relevant 112 16

(e): Human decisions: "N" = excluded, "Y" = included. Machine Learning predictions: "irrelevant" or "relevant"

Unfortunately, the improved performance that was observed during the training phase was not observed on the new dataset. A possible explanation could be that the new data were not fully in line with the original data or possibly, some over-fitting has occurred. Either way, it can be recommended to update the models that were trained, using manual feedback from a reviewer. This feedback would consist of adding all abstracts that were labelled to be relevant by the machine, but assign the correct value based on the reviewer's expert knowledge.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

4. Conclusions

Systematic reviews aim at answering specific research questions through the collection and critical analysis of relevant evidence, for example, from multiple research studies or papers. A main advantage of SRs is that it is often cheaper to review existing studies, rather than performing a new study. Nevertheless, conducting a SR can be highly time-consuming, hence the need for automation.

The paper by Tsafnat et al. (2014) presents an overview of the different steps that are required to implement an appropriate SR. Since it is current practice at EFSA, one additional step, called critical appraisal, was discussed as well. For each of these steps, this report presented possible ways for automation. Major interest was in automation through machine learning techniques, which encompasses the construction and use of algorithms that can learn from and make predictions on data. It was noted that the use of MLTs was not possible for all of the steps.

The most important step that could benefit from these MLTs was the screening of abstracts step. Indeed, in a SR, the initial pool of possibly relevant articles can be quite large and determining which of these articles should be included requires a lot of human effort by the reviewers. The time that is allocated to this step can be greatly reduced if one could train an algorithm that allows the automatic determination of relevance of abstracts. In this respect, the report introduced several MLTs that can accomplish this task. Initially, the abstracts should be converted into an appropriate input format that can be passed on to the algorithms. A standard and often-used input format is the bag-of-words approach, which creates a term-document matrix, indicating for each of the documents, the frequency of the individual words in the pool of abstracts. Underlying connections between the individual words can be accounted for by the topic modelling approach (using latent Dirichlet allocation) or by considering n-grams (using the frequency of the combinations of n-words) as input space. In a second step, specific MLT based classifiers can be trained with the purpose of assigning the labels relevant or irrelevant to the abstracts. This training is performed on only a part of the total pool of abstracts (called the training set), which comprises a set of manually labelled data. Once the classifier is optimized, the remaining part of the pool of abstracts can be classified automatically.

A discussion on the strengths and weaknesses of the introduced MLTs was provided. In this respect, attention was paid to the general, intrinsic properties of the different methods as well as to their classification performance, specifically in the field of text classification. In spite of their rather complex nature, SVMs were found to be very popular. Not only are they able to cope with a high number of features and irrelevant features, they are also able to create both linear and non-linear decision boundaries, thereby making the classification rule more general. In addition, they are easier to interpret compared to neural networks, which are generally considered to be black-box algorithms. An even more intuitive interpretation is obtained by using the method of boosting classification trees (i.e. combining many weak classifiers to produce a powerful committee) or by considering random forests (i.e. combining fully grown decision trees to make a final decision). Based on their intrinsic properties, these four different MLTs were considered to be the most interesting for implementation purposes. After implementation, the MLTs were tested on three different case studies and the corresponding performance measures were compared. In this way, a more formal comparison in the field of abstract screening was achieved.

From all case studies that were discussed in this report, similar conclusions could be derived. Most importantly, it was noted that all classifiers that were selected were trained using either the SMOTE or the ROSE correction for imbalanced data. Using these corrections, it was observed that especially the random forests and neural networks were good individual classifiers in terms of high sensitivity with appropriate specificity. This means that these classifiers are able to achieve a high percentage of correctly identified relevant papers, while only leaving an acceptable amount of irrelevant papers in the selected pool of relevant articles. In terms of the feature space, it was seen that for the topics approach,

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

typically less data were required in the training set to obtain adequate classifiers. This means less manual effort for the reviewers in attaching labels to the abstracts in the training set to be used in constructing the classifiers. Moreover, another important advantage of the topics feature space over the TDM feature space is the computation time. Indeed, 30 topics were found to be sufficient in achieving suitable classifiers, while the TDM consisted of over 300 terms in the Isoflavones case study. The more variables that are used in the training process, the longer the computation time. Hence, although the construction of the topics takes longer than constructing the TDM (approximately 15 minutes for the largest case study), it is clear that the topics approach is favoured in terms of computation time as well, especially when it is kept in mind that larger training sets are required to achieve good classifiers with the TDM approach.

While the mentioned individual classifiers already gave good results, better results could be observed when ensembles were created. In this perspective, it is important to note that also classifiers that were trained on the original training sample (without correction for imbalance) were occasionally selected in the optimal ensemble. This shows that the selection of MLTs to be included in the ensemble is not straightforward and no general guidelines can be presented to lead the user towards an optimal solution. Indeed, it was observed that it is not only important which type of classifier is included, but also in which order they appear in the ensemble. Typically, it was observed that classifiers with a high sensitivity should be combined with classifiers with a high specificity. Nevertheless, this is not a general rule and the user is required to try out which combinations provide the best results when applying this approach in a new systematic review. This trial-and-error process might take around 10 minutes on the average.

Although the main focus in the results section was addressed to the screening of abstracts, some attention was paid to the full text screening and prediction of new data as well. In the light of comparing full text screening to abstract screening, it was observed that the amount of correctly identified relevant papers was somewhat smaller in full text screening than in the abstract screening part, but the amount of correctly identified irrelevant articles was higher. As such, there was no clear winner as to which method is preferred. However, this could be due to the fact that only a limited amount of training data was available and a more in-depth study could be recommended for the future. Predicting new data was performed for the ERIS case study, where 318 new abstracts were classified. It could be observed that the selected models were able to identify at least 65% of the relevant abstracts. This also means that there still are some relevant abstracts which could not be identified. This is a threat that the user should be aware of. Nevertheless, these were only initial explorations, using models that were trained on a dataset with only a limited amount of information.

In addition to the screening of abstracts, the MLTs introduced in this report could also be employed for the purpose of data extraction and in relation to the critical appraisal of included studies. Nevertheless, it is important to note here that the current applications of the MLTs in the latter settings are not well studied yet. To date, the applications found in literature are rather specific to the domains of interest in those papers, so a modification of the methodology is required. A more general method was provided by Basu et al. (2016), employing support vector machines on selected sentences from the articles. The selection of these sentences is less straightforward compared to for instance the construction of the term-document matrix, as it requires the application of techniques that are comprised under natural language processing. Similar techniques are employed in the light of the critical appraisal of studies, where Marshall et al. (2015) use support vector machines to identify possible bias in the studies and Millard et al. (2016) employs logistic regression models for the same task. The implementation of these MLTs were considered to be out of the scope of the current report, mainly due to the fact that these methods are not studied in full detail yet, but also due to the lack of available data. Indeed, especially for the critical appraisal stage, information is not only required on the article level, but also on the sentence-level. Both levels of information were lacking in the considered case studies.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Although creating a universal automatic data extraction tool was considered to be infeasible in this stage, this step of the systematic review was addressed in a more basic manner. More specifically, the R shiny tool that was constructed to aid in the screening for relevance also allows for searching the pool of full texts for specific, user-defined words and to show the context in which these words appear. In this way, the user can more efficiently retrieve specific data elements from the papers.

For the steps in the SR process that could not benefit from MLTs, more general ways for automation were presented as well. For example, with respect to the search for articles, automatic query expansion algorithms are used by most existing databases. These algorithms use techniques such as synonym expansion and word sense disambiguation. In addition, web-crawling software using Application Program Interfaces (APIs) have also proved to be useful to automate the search procedure. Once the articles have been retrieved (possibly using specific R packages), duplicates can be automatically identified by reference managers. Bramer et al. (2016) presents a new method that optimizes the de-duplication using Endnote. As mentioned above, the stage of data extraction highly depends on natural language processing techniques. In order to create a useful data extraction tool, there should be a high symbiosis between NLP and MLTs. Once data have been extracted and appraised, summaries should be created, either narrative or quantitative. For the latter, several R packages exist to aid in performing meta-analyses.

In summary, the authors believe that the results presented in this report provide proof that the developed shiny application could be efficiently used in combination with the DistillerSR tool from EFSA, where the latter can be used to create the input files in the correct format and the application could be utilized to reduce the manual workload for the reviewers in the steps of screening the abstracts and full texts. Also searching for important key words in the pool of relevant articles can be done in a simple, yet fast manner using the application.

5. Recommendations

Based on the above-mentioned strengths and weaknesses, it can be concluded the there is definitely an opportunity to use the introduced MLTs for automating the screening of abstracts and full texts steps, at least partially. Indeed, some manual effort is still required to label training data from which the classifiers can learn to predict the correct outcomes. Nevertheless, the reviewers need to be aware that the constructed classifiers are not perfect. Relevant papers can be predicted to be irrelevant and vice versa, thereby resulting into two possible errors. When training the classifiers, one should aim at minimising the former error by striving for a high sensitivity. Unfortunately, as observed by considering the distinct validation datasets, applying classifiers to new data might lead to a reduction of the sensitivity observed on the test set. To keep this threat to a minimum, it can be advised (at least in an initial phase of implementation) to run the automated classifier in parallel with a human reviewer, compare the abstracts that gave different relevant indicators, adjust the indicator where necessary and retrain the classifiers on the revised pool of abstracts. This leads to the following workflow:

- Step 1: A human reviewer assigns an indicator of relevance to a subset of the abstracts under consideration. Based on the QPS case study, an adequate classifier could already be constructed on 50% of the training data (i.e. around 1500 abstracts). Hence, when a labelled dataset of approximately this magnitude is available, the shiny application can be used to train the classifiers.
- Step 2: Use the shiny application to train all the classifiers. The user can select the most appropriate (ensemble of) classifier(s) and predict the remaining abstracts that were not labelled yet.
- Step 3: A detailed investigation of (a subset of) the abstracts that were classified as relevant should be performed by a human reviewer. After the reviewer has adjusted the assigned label

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

when necessary, the revised dataset can be used again to train updated classifiers with the shiny application.

- Step 4: After the new classifier or ensemble of classifiers have been selected, a new prediction of the abstracts can be made.
- Step 5: The full texts of the relevant abstracts could then be retrieved (manually) and the pdf files can be uploaded to the shiny application to train classifiers for the full text screening. For this purpose, Steps 1-4 could be repeated.

It should be noted that the investigated case studies contained less than 7000 abstracts. Nevertheless, much larger datasets are not an exception when doing a systematic review. It is difficult to predict how the classifiers would react on bigger datasets, both in terms of performance and computation time. In this respect, it can be advised to follow steps 1-4 above, also using a small subset of the data (magnitude specified above, i.e. around 1500 abstracts). In this setting, it is important that the used subset contains enough information on both relevant and irrelevant papers. Based on the QPS case study, where the classifiers could handle the class imbalance of having only 4% relevant papers, around 60 relevant abstracts would suffice to train the classifiers. Once the classifiers are constructed, the remaining abstracts can be classified automatically. Again here, it is important to note that classification errors are not avoidable and the amount of errors depends on the quality of the training set.

In case of a systematic review that is recurring after for example a year, steps 1-5 mentioned above are only required in the starting phase. Once enough data have been collected, the trained classifiers could be used for new data that enter over the next year. Our initial exploration in Section 3.2.2 already indicated a promising future for the developed tool, but the current dataset is not detailed enough at this time to provide a high-performing classifier. Of course, of time allows, it is always recommended to update the classifiers each year to include more information and increase the performance of the MLTs.

Over the course of the project, several interesting ideas were raised during the interim meetings. These ideas were mostly related to the generality of the proposed methods. For example, the classifiers that are trained for one specific systematic review cannot be used for determining the relevance of abstracts in another, unrelated systematic review. In this respect, the use of deep learning, e.g. through multilayer neural networks, could provide a solution. More specifically, some of the hidden layers of the constructed neural networks could answer more general question and possibly be re-used in different reviews. As a second alternative, the standard MLTs could be used to answer more general questions that underlie the process of manual screening of abstracts. As such, these classifiers could also be recycled across multiple SRs. Another idea for further improvement of the automation process is to develop the MLT methodology for the purpose of data extraction and critical appraisal. However, making universal tools for this latter purpose would be the subject of a major research application.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

References

Aljaber B, Stokes N, Bailey J and Pei J 2010. Document clustering of scientific texts using citation contexts. Information Retrieval, 13, 101–131.

Basu T, Kumar S, Kalyan A, Jayaswal P, Goyal P, Pettifer S and Jonnalagadda SR 2016. A Novel Framework to Expedite Systematic Reviews by Automatically Building Information Extraction Training Corpora. arXiv:1606.06424v1

Bekhuis T, and Demner-Fushman D 2010. Towards automating the initial screening phase of a systematic review. Studies in Health Technology and Informatics, 160, 146–150.

Blei DM, Ng AY and Jordan MI 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022.

Boudin F, Nie J-Y, Bartlett JC, Grad R, Pluye P and Dawes M 2010a. Combining Classifiers for Robust PICO Element Detection. BMC Medical Informatics and Decision Making, 10 (29), 1–6.

Boudin F, Nie J-Y and Dawes M 2010b. Clinical Information Retrieval using Document and PICO Structure. Proceedings of the Human Language Technologies–North American Association of Computational Linguistics, 822–830.

Bramer WM, Holland L, Mollema J, Hannon T and Bekhuis T 2014. Removing duplicates in retrieval sets from electronic databases: comparing the efficiency and accuracy of the Bramer-method with other methods and software packages. Proceedings of the 14th European Association for Health Information and Libraries (EAHIL) Conference.

Bramer WM, Giustini D, de Jonge GB, Holland L and Bekhuis T 2016. De-duplication of database search results for systematic reviews in EndNote. Journal of the Medical Library Association, 104 (3),140–143.

Brodley C E, Rebbapragada U, Small K and Wallace BC 2012. Challenges and Opportunities in Applied Machine Learning. Artificial Intelligence Magazine, 33 (1), 11–24.

Bui DDA, Del Fiol G and Hurdle JF 2016. Extractive text summarization system to aid data extraction from full text in systematic review development. Journal of Biomedical Informatics, 64, 265–272.

Caropreso, MF, Matwin S and Sebastiani F 2001. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. Text Databases and Document Management: Theory and Practice, A. G. Chin, ed. Idea Group Publishing, Hershey, PA, 78–102.

Choong MK, Galgani F, Dunn AG and Tsafnat G 2014. Automatic Evidence Retrieval for Systematic Reviews. Journal of Medical Internet Research, 16 (10), e223.

Cochrane Collaboration (2013). Template for the protocol of systematic reviews. Available online: http://endoc.cochrane.org/sites/endoc.cochrane.org/files/public/uploads/CMED_protocol_template.pdf

Cohen AM, Hersh WR, Peterson K and Yen PY 2006. Reducing workload in systematic review preparation using automated citation classification. Journal of the American Medical Informatics Association, 13 (2), 206–219.

Councill IG, Giles CL and Kan MY 2008. ParsCit: an open-source CRF reference string parsing package. Proceedings of the Sixth International Language Resources and Evaluation, Presented at: International Language Resources and Evaluation Conference, Marrakesh, Morrocco.

Cortes C and Vapnik V 1995. Support-Vector Networks. Machine Learning, 20, 273–297.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Consoli S and Stilianakis NI 2017. A quartet method based on variable neighborhood search for biomedical literature extraction and clustering. International transactions in operational research, 24 (3), 537–558.

Dalal SR, Shekelle PG, Hempel S, Newberry SJ, Motala A and Shetty KD 2012. A Pilot Study Using Machine Learning and Domain Knowledge to Facilitate Comparative Effectiveness Review Updating. Medical Decision Making, 1–13.

EFSA (European Food Safety Authority), 2010. Application of systematic review methodology to food and feed safety assessments to support decision making. EFSA Journal 2010; 8(6):1637, 90 pp. doi:10.2903/j.efsa.2010.1637.

EFSA ANS Panel (EFSA Panel on Food Additives and Nutrient Sources added to Food), 2015. Scientific opinion on the risk assessment for peri- and post-menopausal women taking food supplements containing isolated isoflavones. EFSA Journal 2015; 13(10):4246, 342 pp. doi:10.2903/j.efsa.2015.4246.

Frunza O, Inkpen D and Matwin S 2010. Building systematic reviews using automatic text classification techniques. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 301–311.

Galavotti L, Sebastiani F and Simi M 2000. Experiments on the use of feature selection and negative evidence in automated text categorization. Proceedings of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries (Lisbon, PT, 2000), 59–68.

Glass GV 1976. Primary, Secondary, and Meta-Analysis of Research. Educational Researcher, 5 (10), 3-8.

Hamad Z and Salim N 2014. Systematic Literature Review (SLR) Automation: A Systematic Literature Review. Journal of Theoretical and Applied Information Technology, 59 (3), 661–672.

Hartling L, Bond K, Vandermeer B, Seida J, Dryden DM and Rowe BH 2011. Applying the risk of bias tool in a systematic review of combination long-acting beta-agonists and inhaled corticosteroids for persistent asthma. PLoS One, 6, e17242.

Hashimoto K, Kontonatsios G, Miwac M and Ananiadou S 2016. Topic detection using paragraph vectors to support active learning in systematic reviews. Journal of Biomedical Informatics, 62, 59–65.

Hastie T, Tibshirani R, and Friedman J 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. Available online: http://web.stanford.edu/~hastie/ElemStatLearn/

Higgins J, Churchill R, Tovey D, Lasserson T and Chandler J 2011. Update on the MECIR project: methodological expectations for Cochrane intervention. Cochrane Methodology, 1, 2–45.

Higgins J and Green S 2011. Cochrane handbook for systematic reviews of interventions version 5.1.0. The Cochrane Collaboration.

Howard BE, Phillips J, Miller K, Tandon A, Mav D, Shah MR, Holmgren S, Pelch KE, Walker V, Rooney AA, Macleod M, Shah RR and Thayer K 2016. SWIFT-Review: a text-mining workbench for systematic review. Systematic Reviews, 5, 1–16.

Ji X and Yen P-Y 2015. Using MEDLINE Elemental Similarity to Assist in the Article Screening Process for Systematic Reviews. Journal of Medical Internet Research, 3 (3), 1–13.

Jiang L and Zhang H 2006. Weightily Averaged One-Dependence Estimators. In: Yang Q., Webb G. (eds) PRICAI 2006: Trends in Artificial Intelligence. PRICAI 2006. Lecture Notes in Computer Science, 4099.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Jonnalagadda S and Petitti D 2013. A New Iterative Method to Reduce Workload in the Systematic Review Process. International journal of computational biology and drug design. 6, 5–17.

Jonnalagadda S, Goyal P and Huffman M 2015. Automating data extraction in systematic reviews: a systematic review. Systematic Reviews, 4, 1–16.

Kaufman L, and Rousseeuw PJ 1990. Finding groups in data: An introduction to cluster analysis. New Jersey: John Wiley and Sons, Inc.

Kate K, Chaudhari S, Prapanca A and Kalagnanam J 2014. FoodSIS: A Text Mining System to Improve the State of Food Safety in Singapore. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1709–1718.

Khabsa M, Elmagarmid A, Ilyas I, Hammady H and Ouzzani M 2016. Learning to identify relevant studies for systematic reviews using random forest and external information. Machine Learning, 102, 465–482.

Kiritchenko S, de Bruijn B, Carini S, Martin J and Sim I 2010. ExaCT: automatic extraction of clinical trial characteristics from journal publications. BMC Medical Informatics and Decision Making, 10 (56), 1–17.

Kouznetsov A and Japkowicz N 2010. Using classifier performance visualization to improve collective ranking techniques for biomedical abstracts classification. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6085 LNAI, 299–303.

Kuper H, Nicholson A and Hemingway H 2006. Searching for observational studies: what does citation tracking add to PubMed? A case study in depression and coronary heart disease. BMC Medical Research Methodology, 6, 4.

Kwon Y, Lemieux M, McTavish J and Wathen N 2015. Identifying and removing duplicate records from systematic review searches. Journal of Medical Library Association, 103 (4), 184–188.

Lafferty JD, McCallum A and Pereira FCN 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labelling Sequence Data. ICML 2001 Proceedings of the Eighteenth International Conference on Machine Learning, 282–289.

Leaman R and Gonzalez G 2008. BANNER: an executable survey of advances in biomedical named entity recognition. Pacific Symposium on Biocomputing, 13, 352–663.

Lensen S, Farquhar C and Jordon V 2014. Risk of bias: are judgements consistent between reviews? Cochrane Database Systematic Reviews, 1, 1–30.

Lin J-W, Chang C-H, Lin M-W, Ebell MH, Chiang J-H 2011. Automating the process of critical appraisal and assessing the strength of evidence with information extraction technology. Journal of Evaluation in Clinical Practice, 17 (4), 832–838.

Liu J, Timsina P and El-Gayar O 2016. A comparative analysis of semi-supervised learning: The case of article selection for medical systematic reviews. Information Systems Frontiers, 1–11.

Lunardon N, Menardi G and Torelli N 2014. ROSE: a Package for Binary Imbalanced Learning, R Journal, 6 (1), 82–91.

Marshall IJ, Kuiper J and Wallace BC 2015. Automating risk of bias assessment for clinical trials. IEEE Journal of Biomedical and Health Informatics, 19, 1406–1412.

Millard LA, Flach PA and Higgins JP 2016. Machine learning to assist risk-of-bias assessments in systematic reviews. International Journal of Epidemiology, 45, 266–77.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Mishra R, Bian J, Fiszman M, Weir CR, Jonnalagadda S, Mostafa J, Del Fiol G 2014. Text summarization in the biomedical domain: A systematic review of recent research. Journal of Biomedical Informatics, 52, 457–467.

Mitchell TM 1997. Machine Learning, McGraw-Hill Science/Engineering/Math, New York, 400pp.

Mikolov T, Sutskever I, Chen K, Corrado GS and Dean J 2013. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems. 3111–3119.

Mo Y, Kontonatsios G and Ananiadou S 2015. Supporting systematic reviews using LDA-based document representations. Systematic Reviews, 4 (172), 1–12.

Ng HT, Goh WB and Low KL 1997. Feature selection, perceptron learning, and a usability case study for text categorization. Proceedings of SIGIR-97, 20th ACM International Conference on Research and Development in Information Retrieval (Philadelphia, US, 1997), 67–73.

Olorisade BK, Quincey E, Brereton P and Andras P 2016. A Critical Analysis of Studies that Address the Use of Text Mining for Citation Screening in Systematic Reviews. Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering.

O'Mara-Eves A, Thomas J, McNaught J, Miwa M and Ananiadou S 2015. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Systematic reviews, 4(1), 1–22.

Paek H, Kogan Y, Thomas P, Codish S and Krauthammer M 2006. Shallow semantic parsing of randomized controlled trial reports. AMIA 2006 Symposium Proceedings, 604–608.

Qi X, Yang M, Ren W, Jia J, Wang J, Han G and Fan D 2013. Find duplicates among the PubMed, EMBASE, and Cochrane Library Databases in systematic review. PLOS One, 8 (8), e71838.

Rathbone J, Carter M, Hoffmann TI and Glasziou P 2015. Better duplicate detection for systematic reviewers: evaluation of systematic review assistant-deduplication module. Systematic Reviews, 4(1), 1–6.

Rennie J, Shih L, Teevan J and Karger D 2003. Tackling the poor assumptions of naive Bayes text classifiers, Proceedings of the twentieth international conference on machine learning (ICML), Washington, DC.

Robinson DA 2012. Finding patient-oriented evidence in PubMed abstracts. Athens: University of Georgia, 1-66.

Rosario B and Hearst MA 2005. Multi-way relation classification: application to protein-protein interactions. Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, Columbia, 732–739.

Rodriguez-Esteban R and Iossifov I 2009. Figure mining for biomedical research. Bioinformatics, 25(16), 2082–2084.

Ru G, Crescio I, Gregori D et al., 2017. Machine Learning Techniques applied in risk assessment related to food safety. EFSA Supporting Publication 2017:EN-1254, 311pp. doi:10.2903/sp.efsa.2017.EN-1254.

Saha TK, Ouzzani M, Hammady HM and Elmagarmid AK 2016. A large scale study of SVM based methods for abstract screening in systematic reviews. arXiv:1610.00192.

Sebastiani F 2002. Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34(1), 1–47.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Sellak H, Ouhbi B and Frikh B 2015. Using Rule-based Classifiers in Systematic Reviews: A Semantic Class Association Rules Approach. Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services, Article No. 43.

Shemilt I, Simon A, Hollands GJ, Marteau TM, Ogilvie D, O'Mara-Eves A, Kelly MP and Thomas J 2013. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. Research Synthesis Methods, 5(1), 31–49.

Stansfield C, Thomas J and Kavanagh J 2013. 'Clustering' documents automatically to support scoping reviews of research: a case study. Research synthesis methods, 4 (3), 230–241.

Summerscales RL, Argamon S, Bai S, Hupert J and Shwartz A 2011. Automatic summarization of results from clinical trials. Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference, 372–377.

Sutton A, Abrams K, Jones D, Sheldon T, and Song F 2000. Methods for Meta-Analysis in Medical Research. John Wiley & Sons, Ltd, Chichester, UK.

Thomas J, McNaught J and Ananiadou S 2011. Applications of text mining within systematic reviews. Res Synth Meth, 2, 1–14.

Timsina P, Liu J and El-Gayar O 2016. Advanced analytics for the automation of medical systematic reviews. Information Systems Frontiers, 18 (2), 237–252.

Tsafnat G, Glasziou P, Choong MK, Dunn A, Galgani F and Coiera E 2014. Systematic review automation technologies. Systematic Reviews, 3, 1–15.

Tsertsvadze A, Chen Y-F, Moher D, Sutcliffe P and McCarthy N 2015. How to conduct systematic reviews more expeditiously? Systematic Reviews, 4 (160), 1–6.

Wallace BC, Trikalinos TA, Lau J, Brodley C and Schmid CH 2010. Semi-automated screening of biomedical citations for systematic reviews. BMC bioinformatics, 11, 55.

Wallace BC, Kuiper J, Sharma A, Zhu M and Marshall IJ 2016. Extracting PICO Sentences from Clinical Trial Reports using Supervised Distant Supervision. Journal of Machine Learning Research, 17, 1–25.

Wang J, Shen X and Pan W 2007. On transductive support vector machines. Prediction and Discovery, American Mathematical Society, 443, 7–19.

Webb GI, Boughton JR and Wang Z 2005. Not So Naive Bayes: Aggregating One-Dependence Estimators. Machine Learning, 58(1), 5–24.

Yang Y and Pedersen J 1997. A comparative study on feature selection in text categorization. In Proceedings of ICML-97, the 14th International Conference on Machine Learning.

Yang H, Huang K, King I and Lyu MR 2015. Maximum margin semi-supervised learning with irrelevant data. Neural Networks, 70, 90–102.

Yu Z, Bernstam E, Cohen T, Wallace BC, Johnson TR 2016. Improving the utility of MeSH® terms using the TopicalMeSH representation. Journal of Biomedical Informatics, 61, 77–86.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Abbreviations

(W)AODE	(Weightily) Averaged One-Dependence Estimator
BOW	Bag of Words
CAR	Class Association Rule
CAT	Critical Appraisal Tool
CNB	Complement Naïve Bayes
EFSA	European Food Safety Authority
ERIS	Emerging Risk Identification System
GBM	Gradient Boosting Machine
GLM	Generalized Linear Models
LDA	Latent Dirichlet Allocation
MeSH	Medical Subject Headings
ML	Machine Learning
MLT	Machine Learning Technique
NER	Named Entity Recognition
NLP	Natural Language Processing
NN	Neural Network
PICO	Patient, Intervention, Comparison, and Outcome
QPS	Qualified Presumption of Safety
RF	Random Forest
ROSE	Random Over-Sampling Examples
SCER	Scientific Committee and Emerging Risk (EFSA Unit)
SMOTE	Synthetic Minority Over-sampling Technique
SR	Systematic Review
SRA-DM	Systematic Review Assistant-Deduplication Module
SVM	Support Vector Machine
SWOT	Strengths, Weaknesses, Opportunities, and Threats
TDM	Term-Document Matrix
TF-IDF	Term Frequency-Inverse Document Frequency
TM	Text Mining
TP/TN/FP/FN	True Positive/True Negative/False Positive/False Negative
UMLS	Unified Medical Language System

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Appendix A – Performance measures from case studies

In this appendix, the individual performance measures as obtained from the R shiny application are presented for all three case studies. Figure 3: - Figure 8: pertain to the Isoflavones case study, Figure 3: - Figure 14: are related to the QPS case study and Figure 15: - Figure 20: show the results of the ERIS case study.

Show 25 entries Search:

	Accuracy	Kappa	F1	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
svm_Linear_orig	0.946290395994538	0.146398602607434	0.163120567375887	0.100436681222707	0.992797118847539	0.433962264150942	0.952545496429394	0.433962264150943	0.100436681222707
svm_Linear_smote	0.906463359126081	0.322914737378047	0.368663594470046	0.524017467248908	0.927490996398559	0.28436018957346	0.972557905337362	0.28436018957346	0.524017467248908
svm_Linear_rose	0.887118798361402	0.303193050551039	0.354166666666667	0.593886462882096	0.903241296518607	0.252319109461967	0.975875486381323	0.252319109461967	0.593886462882096
svm_Poly_orig	0.947655894401456	0.078535387151537	0.0873015873015873	0.0480349344978166	0.997118847539016	0.478260869565216	0.950125829329673	0.478260869565217	0.0480349344978166
svm_Poly_smote	0.89804278561675	0.224179748002821	0.275080906148867	0.37117903930131	0.927010804321729	0.218508997429306	0.964044943820225	0.218508997429306	0.37117903930131
svm_Poly_rose	0.849112426035503	0.196928990556296	0.259217877094972	0.506550218340611	0.867947178871549	0.174174174174174	0.969688841201717	0.174174174174174	0.506550218340611
svm_Radial_orig	0.947883477469276	0		0	1		0.947883477469276		0
svm_Radial_smote	0.947883477469276	0		0	1		0.947883477469276		0
svm_Radial_rose	0.947883477469276	0		0	1		0.947883477469276		0
GBM_orig	0.950842057350933	0.304380966899784	0.325	0.22707423580786	0.990636254501801	0.571428571428571	0.958865907506391	0.571428571428571	0.22707423580786
GBM_smote	0.917159763313609	0.384912391080127	0.425867507886435	0.589519650655022	0.935174069627851	0.333333333333333	0.976435196791176	0.333333333333333	0.589519650655022
GBM_rose	0.0609922621756941	0.00098426748060347	0.099912739965096	1	0.00936374549819928	0.0525832376578645	1	0.0525832376578645	1
NN_orig	0.935821574874829	0.304504108063568	0.338028169014084	0.314410480349345	0.969987995198079	0.365482233502538	0.962592327853228	0.365482233502538	0.314410480349345
NN_smote	0.837278106508876	0.220240184220989	0.282848545636911	0.615720524017467	0.849459783913565	0.18359375	0.97573083287369	0.18359375	0.615720524017467
NN_rose	0.844105598543468	0.236549463003255	0.297435897435897	0.633187772925764	0.855702280912365	0.194369973190349	0.976973684210526	0.194369973190349	0.633187772925764
RF_orig	0.950159308147474	0.244598759374948	0.262626262626263	0.170305676855895	0.993037214885954	0.573529411764707	0.956079519186315	0.573529411764706	0.170305676855895
RF_smote	0.909421939007738	0.385160272794584	0.42816091954023	0.650655021834061	0.923649459783914	0.319057815845824	0.979628214922333	0.319057815845824	0.650655021834061
RF_rose	0.204597177969959	0.0176154630619146	0.114068441064639	0.982532751091703	0.161824729891957	0.0605489773950484	0.994100294985251	0.0605489773950484	0.982532751091703

Showing 1 to 18 of 18 entries Previous 1 Next

Figure 3: Performance measures of the individual classifiers for the Isoflavones case study using 20% of training data and TDM as input space

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Show 25 entries Search:

	Accuracy	Kappa	F1	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
svm_Linear_orig	0.947560087399854	0.110082864846133	0.121951219512195	0.0699300699300699	0.995774106799846	0.476190476190477	0.951192660550459	0.476190476190476	0.0699300699300699
svm_Linear_smote	0.892935178441369	0.359804227878495	0.407258064516129	0.706293706293706	0.90318862850557	0.286118980169972	0.982448809026327	0.286118980169972	0.706293706293706
svm_Linear_rose	0.857611070648216	0.281699111160469	0.338409475465313	0.699300699300699	0.866308106031502	0.223214285714286	0.981288076588338	0.223214285714286	0.699300699300699
svm_Poly_orig	0.943554260742899	0.182420985097908	0.205128205128205	0.13986013986014	0.987706492508644	0.384615384615384	0.9543429844098	0.384615384615385	0.13986013986014
svm_Poly_smote	0.884195193008012	0.293766549288371	0.345679012345679	0.587412587412587	0.900499423741836	0.244897959183673	0.975447357469829	0.244897959183673	0.587412587412587
svm_Poly_rose	0.922796795338674	0.371664051082678	0.411111111111111	0.517482517482518	0.945063388398002	0.341013824884793	0.972716488730724	0.341013824884793	0.517482517482518
svm_Radial_orig	0.947924253459578	0		0	1		0.947924253459578		0
svm_Radial_smote	0.0524399126001457	0.00004002587406852 76	0.0990304709141274	1	0.000384172109104879	0.0520947176684882	1	0.0520947176684882	1
svm_Radial_rose	0.947924253459578	0		0	1		0.947924253459578		0
GBM_orig	0.947195921340131	0.32453798106121	0.349775784753363	0.272727272727273	0.9842489435267	0.487500000000001	0.96099024756189	0.4875	0.272727272727273
GBM_smote	0.92680262199563	0.441349071556825	0.477922077922078	0.643356643356643	0.942374183634268	0.380165289256198	0.979632587859425	0.380165289256198	0.643356643356643
GBM_rose	0.20721048798252	0.0175963720013318	0.113960113960114	0.979020979020979	0.164809834805993	0.0605012964563526	0.993055555555556	0.0605012964563526	0.979020979020979
NN_orig	0.927166788055353	0.359621649732165	0.397590361445783	0.461538461538462	0.9527468305801	0.349206349206349	0.969886585842785	0.349206349206349	0.461538461538462
NN_smote	0.856518572469046	0.274994672645812	0.332203389830508	0.685314685314685	0.865923933922397	0.21923937360179	0.980426272292301	0.21923937360179	0.685314685314685
NN_rose	0.834304442825929	0.269867779934761	0.329896907216495	0.783216783216783	0.837111025739531	0.208955223880597	0.985972850678733	0.208955223880597	0.783216783216783
RF_orig	0.945010924981792	0.272292988041377	0.297674418604651	0.223776223776224	0.984633115635805	0.444444444444444	0.958489154824233	0.444444444444444	0.223776223776224
RF_smote	0.935178441369264	0.480707733692569	0.513661202185792	0.657342657342657	0.950441797925471	0.42152466367713	0.980578676179152	0.42152466367713	0.657342657342657
RF_rose	0.387108521485797	0.0519570246808175	0.143511450381679	0.986013986013986	0.354206684594698	0.0773874862788145	0.997835497835498	0.0773874862788145	0.986013986013986

Showing 1 to 18 of 18 entries Previous 1 Next

Figure 4: Performance measures of the individual classifiers for the Isoflavones case study using 50% of training data and TDM as input space

Show 25 entries Search:

	Accuracy	Kappa	F1	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
svm_Linear_orig	0.949908925318761	0.265917876157829	0.285714285714286	0.192982456140351	0.99135446685879	0.549999999999999	0.957328385899815	0.55	0.192982456140351
svm_Linear_smote	0.876138433515483	0.318267729476539	0.37037037037037	0.701754385964912	0.88568683957733	0.251572327044025	0.981895633652822	0.251572327044025	0.701754385964912
svm_Linear_rose	0.825136612021858	0.247220139260846	0.309352517985612	0.754385964912281	0.829010566762728	0.194570135746606	0.984036488027366	0.194570135746606	0.754385964912281
svm_Poly_orig	0.938069216757741	0.162527760952957	0.19047619047619	0.140350877192982	0.981748318924111	0.296296296296296	0.954248366013072	0.296296296296296	0.140350877192982
svm_Poly_smote	0.92896174863388	0.290097975829314	0.327586206896552	0.333333333333333	0.961575408261287	0.322033898305084	0.963426371511068	0.322033898305085	0.333333333333333
svm_Poly_rose	0.891621129326047	0.32114467408585	0.37037037037037	0.614035087719298	0.906820365033622	0.265151515151515	0.977225672877847	0.265151515151515	0.614035087719298
svm_Radial_orig	0.948087431693989	0		0	1		0.948087431693989		0
svm_Radial_smote	0.0519125683060109	0	0.0987012987012987	1	0	0.0519125683060109		0.0519125683060109	1
svm_Radial_rose	0.948087431693989	0		0	1		0.948087431693989		0
GBM_orig	0.953551912568306	0.319305666982714	0.337662337662338	0.228070175438596	0.993275696445725	0.649999999999998	0.959183673469388	0.65	0.228070175438596
GBM_smote	0.931693989071038	0.455162558056449	0.489795918367347	0.631578947368421	0.948126801152738	0.4	0.979166666666667	0.4	0.631578947368421
GBM_rose	0.153005464480874	0.0122403496247802	0.109195402298851	1	0.106628242074928	0.0577507598784195	1	0.0577507598784195	1
NN_orig	0.918943533697632	0.298669422555224	0.340740740740741	0.403508771929825	0.94716618635927	0.294871794871795	0.966666666666667	0.294871794871795	0.403508771929825
NN_smote	0.855191256830601	0.240505337892511	0.299559471365639	0.596491228070175	0.869356388088377	0.2	0.975215517241379	0.2	0.596491228070175
NN_rose	0.824225865209472	0.240631248521854	0.303249097472924	0.736842105263158	0.829010566762728	0.190909090909091	0.982915717539863	0.190909090909091	0.736842105263158
RF_orig	0.95264116575592	0.330000704076602	0.35	0.245614035087719	0.99135446685879	0.608695652173912	0.96	0.608695652173913	0.245614035087719
RF_smote	0.932604735883424	0.451718693064496	0.486111111111111	0.614035087719298	0.950048030739673	0.402298850574713	0.978239366963403	0.402298850574713	0.614035087719298
RF_rose	0.467213114754098	0.0695966557451136	0.158273381294964	0.964912280701754	0.439961575408261	0.0862068965517241	0.995652173913044	0.0862068965517241	0.964912280701754

Showing 1 to 18 of 18 entries Previous 1 Next

Figure 5: Performance measures of the individual classifiers for the Isoflavones case study using 80% of training data and TDM as input space

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Show 25 entries Search:

	Accuracy	Kappa	F1	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
svm_Linear_orig	0.947883477469276	0		0	1		0.947883477469276		0
svm_Linear_smote	0.812926718252162	0.231865558799948	0.296232876712329	0.755458515283843	0.81608643457383	0.184238551650692	0.983791606367583	0.184238551650692	0.755458515283843
svm_Linear_rose	0.767410104688211	0.192351755645233	0.262626262626263	0.794759825327511	0.765906362545018	0.157303370786517	0.985480383070744	0.157303370786517	0.794759825327511
svm_Poly_orig	0.946517979062358	0.00493777668991901	0.00843881856540084	0.00436681222707424	0.998319327731092	0.125	0.948016415868673	0.125	0.00436681222707424
svm_Poly_smote	0.88848429676832	0.25132466289685	0.303977272727273	0.467248908296943	0.911644657863145	0.225263157894737	0.968869609594284	0.225263157894737	0.467248908296943
svm_Poly_rose	0.760355029585799	0.191324245126411	0.26208829712684	0.816593886462882	0.757262905162065	0.156093489148581	0.986858573216521	0.156093489148581	0.816593886462882
svm_Radial_orig	0.947883477469276	0		0	1		0.947883477469276		0
svm_Radial_smote	0.947428311333637	0.00680646202921986	0.00858369098712446	0.00436681222707424	0.999279711884754	0.250000000000013	0.948063781321185	0.25	0.00436681222707424
svm_Radial_rose	0.947883477469276	0		0	1		0.947883477469276		0
GBM_orig	0.945607646791079	0.0526772521518892	0.0627450980392157	0.0349344978165939	0.995678271308523	0.307692307692306	0.949404761904762	0.307692307692308	0.0349344978165939
GBM_smote	0.847974510696404	0.244177716572052	0.304166666666667	0.637554585152838	0.859543817527011	0.199726402188783	0.977340977340977	0.199726402188782	0.637554585152838
GBM_rose	0.647018661811561	0.125360599289774	0.205837173579109	0.877729257641921	0.634333733493397	0.116589327146172	0.989513108614232	0.116589327146172	0.877729257641921
NN_orig	0.947883477469276	0		0	1		0.947883477469276		0
NN_smote	0.837505689576695	0.241341353299095	0.302734375	0.676855895196507	0.846338535414166	0.19496855345912	0.979438732981384	0.19496855345912	0.676855895196507
NN_rose	0.786299499317251	0.211622454614416	0.279355333844973	0.794759825327511	0.785834333733493	0.169459962756052	0.985843373493976	0.169459962756052	0.794759825327511
RF_orig	0.947883477469276	0		0	1		0.947883477469276		0
RF_smote	0.881201638598088	0.257320852099975	0.311345646437995	0.51528384279476	0.901320528211284	0.223062381852552	0.971280724450194	0.223062381852552	0.51528384279476
RF_rose	0.69458352298589	0.154847159394009	0.231386025200458	0.882096069868996	0.684273709483794	0.133157547791694	0.990615224191866	0.133157547791694	0.882096069868996

Showing 1 to 18 of 18 entries Previous 1 Next

Figure 6: Performance measures of the individual classifiers for the Isoflavones case study, using 20% of training data and topics as input space

Show 25 entries Search:

	Accuracy	Kappa	F1	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
svm_Linear_orig	0.947924253459578	0		0	1		0.947924253459578		0
svm_Linear_smote	0.834668608885652	0.261923227663522	0.322388059701492	0.755244755244755	0.839031886285056	0.204933586337761	0.984227129337539	0.204933586337761	0.755244755244755
svm_Linear_rose	0.815003641660597	0.238462302746293	0.302197802197802	0.769230769230769	0.817518248175182	0.188034188034188	0.984729291994447	0.188034188034188	0.769230769230769
svm_Poly_orig	0.947195921340131	0.0887682569034085	0.0993788819875776	0.0559440559440559	0.996158278908951	0.444444444444447	0.950513196480938	0.444444444444444	0.0559440559440559
svm_Poly_smote	0.875819373634377	0.253088911187721	0.308316430020284	0.531468531468531	0.894736842105263	0.217142857142857	0.9720367278798	0.217142857142857	0.531468531468531
svm_Poly_rose	0.792061179898033	0.222346323274866	0.288916562889166	0.811188811188811	0.791010372646946	0.175757575757576	0.98705656759348	0.175757575757576	0.811188811188811
svm_Radial_orig	0.947924253459578	0		0	1		0.947924253459578		0
svm_Radial_smote	0.949380917698471	0.0517343323776841	0.054421768707483	0.027972027972028	1	1	0.949307075127644	1	0.027972027972028
svm_Radial_rose	0.947924253459578	0		0	1		0.947924253459578		0
GBM_orig	0.949016751638747	0.180190131744494	0.195402298850575	0.118881118881119	0.994621590472532	0.548387096774196	0.953591160220995	0.548387096774194	0.118881118881119
GBM_smote	0.88856518572469	0.328433574325177	0.378048780487805	0.65034965034965	0.901651940069151	0.26647564469914	0.979140592407176	0.26647564469914	0.65034965034965
GBM_rose	0.67115804806992	0.145505772468726	0.223559759243336	0.909090909090909	0.658086822896658	0.127450980392157	0.992468134414832	0.127450980392157	0.909090909090909
NN_orig	0.948652585579024	0.0497057220975951	0.0536912751677852	0.027972027972028	0.99923165578179	0.666666666666682	0.949270072992701	0.666666666666667	0.027972027972028
NN_smote	0.840859431900947	0.253692041509579	0.313971742543171	0.699300699300699	0.848636189012678	0.202429149797571	0.980905861456483	0.202429149797571	0.699300699300699
NN_rose	0.762563729060452	0.205998674884629	0.275555555555556	0.867132867132867	0.756819054936612	0.163804491413474	0.990447461035696	0.163804491413474	0.867132867132867
RF_orig	0.949016751638747	0.0507138659383944	0.0540540540540541	0.027972027972028	0.999615827890895	0.799999999999988	0.949288580809923	0.8	0.027972027972028
RF_smote	0.876547705753824	0.308994193702595	0.361581920903955	0.671328671328671	0.887821744141375	0.247422680412371	0.980067854113656	0.247422680412371	0.671328671328671
RF_rose	0.738164603058995	0.182385441763659	0.254922279792746	0.86013986013986	0.73146369573569	0.14963503649635	0.98960498960499	0.14963503649635	0.86013986013986

Showing 1 to 18 of 18 entries Previous 1 Next

Figure 7: Performance measures of the individual classifiers for the Isoflavones case study, using 50% of training data and topics as input space

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Show 25 entries Search:

	Accuracy	Kappa	F1	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
svm_Linear_orig	0.948087431693989	0		0	1		0.948087431693989		0
svm_Linear_smote	0.844262295081967	0.292253760036187	0.349809885931559	0.807017543859649	0.846301633045149	0.223300970873786	0.987668161434977	0.223300970873786	0.807017543859649
svm_Linear_rose	0.759562841530055	0.196447283069707	0.266666666666667	0.842105263157895	0.755043227665706	0.158415841584158	0.988679245283019	0.158415841584158	0.842105263157895
svm_Poly_orig	0.948087431693989	0		0	1		0.948087431693989		0
svm_Poly_smote	0.816939890710383	0.256353233730263	0.31864406779661	0.824561403508772	0.816522574447647	0.197478991596639	0.988372093023256	0.197478991596639	0.824561403508772
svm_Poly_rose	0.777777777777778	0.214467920412364	0.282352941176471	0.842105263157895	0.774255523535062	0.169611307420495	0.988957055214724	0.169611307420495	0.842105263157895
svm_Radial_orig	0.948087431693989	0		0	1		0.948087431693989		0
svm_Radial_smote	0.0664845173041894	-0.00012796519346738	0.098504837291117	0.982456140350877	0.0163304514889529	0.0518518518518519	0.944444444444444	0.0518518518518519	0.982456140350877
svm_Radial_rose	0.948087431693989	0		0	1		0.948087431693989		0
GBM_orig	0.947176684881603	0.132488761749081	0.147058823529412	0.087719298245614	0.994236311239193	0.454545454545455	0.952161913523459	0.454545454545455	0.087719298245614
GBM_smote	0.872495446265938	0.316635992638233	0.369369369369369	0.719298245614035	0.88088376560999	0.248484848484848	0.982851018220793	0.248484848484848	0.719298245614035
GBM_rose	0.674863387978142	0.148082465085347	0.225596529284165	0.912280701754386	0.661863592699328	0.128712871287129	0.992795389048991	0.128712871287129	0.912280701754386
NN_orig	0.948087431693989	0.030486104656575	0.0338983050847458	0.0175438596491228	0.999039385206532	0.499999999999986	0.948905109489051	0.5	0.0175438596491228
NN_smote	0.814207650273224	0.237671274836127	0.301369863013699	0.771929824561403	0.816522574447647	0.187234042553192	0.984936268829664	0.187234042553191	0.771929824561403
NN_rose	0.796903460837887	0.235772205472012	0.300940438871473	0.842105263157895	0.794428434197887	0.183206106870229	0.989234449760766	0.183206106870229	0.842105263157895
RF_orig	0.949908925318761	0.0645041360721252	0.0677966101694915	0.0350877192982456	1	1	0.949817518248175	1	0.0350877192982456
RF_smote	0.89799635701275	0.370728562217924	0.416666666666667	0.701754385964912	0.908741594620557	0.296296296296296	0.982346832814122	0.296296296296296	0.701754385964912
RF_rose	0.700364298724954	0.161076818036145	0.236658932714617	0.894736842105263	0.689721421709894	0.136363636363636	0.99171270718232	0.136363636363636	0.894736842105263

Showing 1 to 18 of 18 entries Previous 1 Next

Figure 8: Performance measures of the individual classifiers for the Isoflavones case study, using 80% of training data and topics as input space

Show 25 entries Search:

	Accuracy	Kappa	F1	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
svm_Linear_orig	0.959772450223486	-0.00236975598919198		0	0.998731501057082	0	0.960943856794142	0	0
svm_Linear_smote	0.908167411621292	0.0857797531773312	0.130769230769231	0.177083333333333	0.93784355179704	0.103658536585366	0.96560731388768	0.103658536585366	0.177083333333333
svm_Linear_rose	0.893945550589191	0.155279821671631	0.201834862385321	0.34375	0.916279069767442	0.142857142857143	0.971748878923767	0.142857142857143	0.34375
svm_Poly_orig	0.960585127996749	-0.000804947070538095		0	0.999577167019027	0	0.960975609756098	0	0
svm_Poly_smote	0.897602600568874	0.0707168180578565	0.118881118881119	0.177083333333333	0.926849894291755	0.0894736842105264	0.965213562307354	0.0894736842105263	0.177083333333333
svm_Poly_rose	0.900446972775295	0.0934480233710269	0.140350877192982	0.208333333333333	0.928541226215645	0.105820105820106	0.966549295774648	0.105820105820106	0.208333333333333
svm_Radial_orig	0.960178789110118	-0.00159472412103374		0	0.999154334038055	0	0.960959739731598	0	0
svm_Radial_smote	0.808614384396587	0.135788837204447	0.192109777015437	0.583333333333333	0.817758985200846	0.114989733059548	0.979736575481256	0.114989733059548	0.583333333333333
svm_Radial_rose	0.960178789110118	-0.00159472412103374		0	0.999154334038055	0	0.960959739731598	0	0
GBM_orig	0.957334416903698	0.0439748271307574	0.0540540540540541	0.03125	0.9949260042283	0.199999999999999	0.961978740801308	0.2	0.03125
GBM_smote	0.828931328728159	0.16844829011502	0.22181146025878	0.625	0.837209302325581	0.134831460674157	0.982142857142857	0.134831460674157	0.625
GBM_rose	0.121495327102804	0.00558817713661803	0.08	0.979166666666667	0.0866807610993657	0.0417036379769299	0.990338164251208	0.0417036379769299	0.979166666666667
NN_orig	0.94676960585128	0.0957103948523628	0.120805369127517	0.09375	0.981395348837209	0.169811320754717	0.963870431893688	0.169811320754717	0.09375
NN_smote	0.885060950833	0.15933322469869	0.207282913165266	0.385416666666667	0.905285412262156	0.14176245210728	0.973181818181818	0.14176245210728	0.385416666666667
NN_rose	0.822836245428687	0.121216836663658	0.177358490566038	0.489583333333333	0.836363636363636	0.108294930875576	0.975826344351258	0.108294930875576	0.489583333333333
RF_orig	0.958553433563592	0.0472969895705235	0.0555555555555556	0.03125	0.996194503171247	0.25	0.962025316455696	0.25	0.03125
RF_smote	0.806989028850061	0.167865540140847	0.222585924713584	0.708333333333333	0.810993657505285	0.132038834951456	0.985611510791367	0.132038834951456	0.708333333333333
RF_rose	0.149532710280374	0.00744004749606076	0.0816147433084686	0.96875	0.116279069767442	0.0426019239578562	0.989208633093525	0.0426019239578562	0.96875

Showing 1 to 18 of 18 entries Previous 1 Next

Figure 9: Performance measures of the individual classifiers for the QPS case study, using 20% of training data and TDM as input space

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Show 25 entries Search:

	Accuracy	Kappa	F1	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
svm_Linear_orig	0.958387516254877	-0.00490035935968545		0	0.997293640054127	0	0.960886571056063	0	0
svm_Linear_smote	0.875812743823147	0.184644250892356	0.232931726907631	0.483333333333333	0.891745602165088	0.153439153439153	0.977020014825797	0.153439153439153	0.483333333333333
svm_Linear_rose	0.862158647594278	0.187711133919941	0.237410071942446	0.55	0.874830852503383	0.151376146788991	0.979545454545455	0.151376146788991	0.55
svm_Poly_orig	0.960988296488947	0		0	1		0.960988296488947		0
svm_Poly_smote	0.853055916775033	0.0821935170417944	0.137404580152672	0.3	0.875507442489851	0.0891089108910891	0.968562874251497	0.0891089108910891	0.3
svm_Poly_rose	0.845903771131339	0.131793691852572	0.185567010309278	0.45	0.861975642760487	0.116883116883117	0.974751338944147	0.116883116883117	0.45
svm_Radial_orig	0.960988296488947	0		0	1		0.960988296488947		0
svm_Radial_smote	0.918725617685306	0.221161714781115	0.260355029585799	0.366666666666667	0.941136671177267	0.201834862385321	0.973407977606718	0.201834862385321	0.366666666666667
svm_Radial_rose	0.960988296488947	0		0	1		0.960988296488947		0
GBM_orig	0.963589076723017	0.231905254521456	0.243243243243243	0.15	0.996617050067659	0.642857142857143	0.966535433070866	0.642857142857143	0.15
GBM_smote	0.858907672301691	0.177272480759655	0.227758007117438	0.533333333333333	0.87212449255751	0.144796380090498	0.978739559605163	0.144796380090498	0.533333333333333
GBM_rose	0.153446033810143	0.00625878373549937	0.0805084745762712	0.95	0.121109607577808	0.0420353982300885	0.983516483516483	0.0420353982300885	0.95
NN_orig	0.935630689206762	0.13458980800491	0.168067226890756	0.166666666666667	0.966847090663058	0.169491525423729	0.966193373901285	0.169491525423729	0.166666666666667
NN_smote	0.85110533159948	0.198976551661398	0.249180327868852	0.633333333333333	0.859945872801083	0.155102040816327	0.982985305491106	0.155102040816327	0.633333333333333
NN_rose	0.810793237971391	0.118553472744676	0.175637393767705	0.516666666666667	0.822733423545332	0.10580204778157	0.976706827309237	0.10580204778157	0.516666666666667
RF_orig	0.961638491547464	0.0868270101640343	0.0923076923076923	0.05	0.998646820027064	0.600000000000004	0.962818003913894	0.6	0.05
RF_smote	0.862808842652796	0.200020707098098	0.249110320284698	0.583333333333333	0.874154262516915	0.158371040723982	0.981017463933181	0.158371040723982	0.583333333333333
RF_rose	0.241222366710013	0.0157990149468815	0.0889929742388759	0.95	0.212449255751015	0.0466830466830467	0.990536277602524	0.0466830466830467	0.95

Showing 1 to 18 of 18 entries Previous 1 Next

Figure 10: Performance measures of the individual classifiers for the QPS case study, using 50% of training data and TDM as input space

Show 25 entries Search:

	Accuracy	Kappa	F1	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
svm_Linear_orig	0.960975609756098	0		0	1		0.960975609756098		0
svm_Linear_smote	0.905691056910569	0.268757687576876	0.30952380952381	0.541666666666667	0.920473773265651	0.216666666666667	0.98018018018018	0.216666666666667	0.541666666666667
svm_Linear_rose	0.808130081300813	0.146917759909719	0.202702702702703	0.625	0.815566835871404	0.120967741935484	0.981670061099796	0.120967741935484	0.625
svm_Poly_orig	0.959349593495935	0.0659741206488062	0.0740740740740741	0.0416666666666667	0.996615905245347	0.333333333333333	0.962418300653595	0.333333333333333	0.0416666666666667
svm_Poly_smote	0.860162601626016	0.209700555854402	0.258620689655172	0.625	0.869712351945855	0.16304347826087	0.982791586998088	0.16304347826087	0.625
svm_Poly_rose	0.930081300813008	0.182206141571574	0.218181818181818	0.25	0.957698815566836	0.193548387096774	0.969178082191781	0.193548387096774	0.25
svm_Radial_orig	0.960975609756098	0		0	1		0.960975609756098		0
svm_Radial_smote	0.904065040650406	0.264696942063348	0.305882352941176	0.541666666666667	0.918781725888325	0.213114754098361	0.98014440433213	0.213114754098361	0.541666666666667
svm_Radial_rose	0.960975609756098	0		0	1		0.960975609756098		0
GBM_orig	0.95609756097561	0.113407015857758	0.129032258064516	0.0833333333333333	0.991539763113367	0.285714285714287	0.963815789473684	0.285714285714286	0.0833333333333333
GBM_smote	0.876422764227642	0.222554890219561	0.269230769230769	0.583333333333333	0.888324873096447	0.175	0.981308411214953	0.175	0.583333333333333
GBM_rose	0.0471544715447154	0.000665502038099995	0.0757097791798107	1	0.00846023688663283	0.039344262295082	1	0.039344262295082	1
NN_orig	0.931707317073171	0.265609007164791	0.3	0.375	0.954314720812183	0.25	0.974093264248705	0.25	0.375
NN_smote	0.847154471544715	0.177971162870062	0.229508196721311	0.583333333333333	0.857868020304569	0.142857142857143	0.980657640232108	0.142857142857143	0.583333333333333
NN_rose	0.796747967479675	0.168370492973745	0.22360248447205	0.75	0.798646362098139	0.131386861313869	0.98744769874477	0.131386861313869	0.75
RF_orig	0.954471544715447	0.10758706467662	0.125	0.0833333333333333	0.998847715736041	0.249999999999999	0.963756177924217	0.25	0.0833333333333333
RF_smote	0.907317073170732	0.33543763862822	0.373626373626374	0.708333333333333	0.915397631133672	0.253731343283582	0.987226277372263	0.253731343283582	0.708333333333333
RF_rose	0.289430894308943	0.0187089919270919	0.0914760914760915	0.916666666666667	0.263959390862944	0.0481400437636762	0.987341772151899	0.0481400437636762	0.916666666666667

Showing 1 to 18 of 18 entries Previous 1 Next

Figure 11: Performance measures of the individual classifiers for the QPS case study, using 80% of training data and TDM as input space

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Show 25 entries Search:

	Accuracy	Kappa	F1	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
svm_Linear_orig	0.960991466883381	0		0	1		0.960991466883381		0
svm_Linear_smote	0.697277529459569	0.10922705213085	0.171301446051168	0.802083333333333	0.693023255813954	0.0958904109589041	0.98854041013269	0.0958904109589041	0.802083333333333
svm_Linear_rose	0.765542462413653	0.134552925552501	0.193006993006993	0.71875	0.767441860465116	0.111470113085622	0.985342019543974	0.111470113085622	0.71875
svm_Poly_orig	0.960991466883381	0		0	1		0.960991466883381		0
svm_Poly_smote	0.885412433969931	0.0563393687554048	0.107594936708861	0.177083333333333	0.914164904862579	0.0772727272727273	0.964747880410531	0.0772727272727273	0.177083333333333
svm_Poly_rose	0.915481511580658	0.06824081051662	0.111111111111111	0.135416666666667	0.947145877378435	0.0942028985507246	0.964270340077486	0.0942028985507246	0.135416666666667
svm_Radial_orig	0.960178789110118	0.016932157239983	0.02	0.0104166666666667	0.998731501057082	0.249999999999995	0.961334961334961	0.25	0.0104166666666667
svm_Radial_smote	0.940268183665177	0.0895622302475099	0.119760479041916	0.104166666666667	0.974207188160677	0.140845070422535	0.964016736401674	0.140845070422535	0.104166666666667
svm_Radial_rose	0.960178789110118	0.016932157239983	0.02	0.0104166666666667	0.998731501057082	0.249999999999995	0.961334961334961	0.25	0.0104166666666667
GBM_orig	0.956115400243803	0.00833513666788491	0.0181818181818182	0.0104166666666667	0.994503171247357	0.0714285714285717	0.961176951369023	0.0714285714285714	0.0104166666666667
GBM_smote	0.843965867533523	0.143234809349655	0.196652719665272	0.489583333333333	0.858350951374207	0.12303664921466	0.976430976430976	0.12303664921466	0.489583333333333
GBM_rose	0.670459162941894	0.0942182793664021	0.157840083073728	0.791666666666667	0.66553911205074	0.0876585928489043	0.987452948557089	0.0876585928489043	0.791666666666667
NN_orig	0.960991466883381	0		0	1		0.960991466883381		0
NN_smote	0.8488419341731	0.123108752705887	0.176991150442478	0.416666666666667	0.866384778012685	0.112359550561798	0.973396674584323	0.112359550561798	0.416666666666667
NN_rose	0.711905729378302	0.11723791459891	0.178447276940904	0.802083333333333	0.708245243128964	0.100391134289439	0.988783943329398	0.100391134289439	0.802083333333333
RF_orig	0.960991466883381	0.0188459845344971	0.0204081632653061	0.0104166666666667	0.999577167019027	0.499999999999971	0.961366409109394	0.5	0.0104166666666667
RF_smote	0.871596911824462	0.160206126786115	0.21	0.4375	0.889217758985201	0.138157894736842	0.974965229485396	0.138157894736842	0.4375
RF_rose	0.734660707029663	0.113708774102958	0.174462705436157	0.71875	0.735306553911205	0.0992805755395683	0.984711211778029	0.0992805755395683	0.71875

Showing 1 to 18 of 18 entries Previous 1 Next

Figure 12: Performance measures of the individual classifiers for the QPS case study, using 20% of training data and topics as input space

Show 25 entries Search:

	Accuracy	Kappa	F1	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
svm_Linear_orig	0.960988296488947	0		0	1		0.960988296488947		0
svm_Linear_smote	0.817945383615085	0.168134755061042	0.222222222222222	0.666666666666667	0.824086603518268	0.133333333333333	0.983844911147011	0.133333333333333	0.666666666666667
svm_Linear_rose	0.750325097529259	0.119602160924475	0.179487179487179	0.7	0.752368064952639	0.102941176470588	0.984070796460177	0.102941176470588	0.7
svm_Poly_orig	0.960338101430429	-0.00128071036735158		0	0.999323410013532	0	0.960962914769031	0	0
svm_Poly_smote	0.879713914174252	0.0798877225660834	0.131455399061033	0.233333333333333	0.90595399188092	0.0915032679738562	0.966787003610108	0.0915032679738562	0.233333333333333
svm_Poly_rose	0.777633289986996	0.119064526501956	0.177884615384615	0.616666666666667	0.784167794316644	0.103932584269663	0.980541455160744	0.103932584269663	0.616666666666667
svm_Radial_orig	0.960988296488947	0		0	1		0.960988296488947		0
svm_Radial_smote	0.874512353706112	0.201256101564476	0.249027237354086	0.533333333333333	0.888362652232747	0.16243654822335	0.979120059656972	0.16243654822335	0.533333333333333
svm_Radial_rose	0.960988296488947	0		0	1		0.960988296488947		0
GBM_orig	0.961638491547464	0.0599995856382243	0.0634920634920635	0.0333333333333333	0.999323410013532	0.666666666666652	0.962214983713355	0.666666666666667	0.0333333333333333
GBM_smote	0.890117035110533	0.242797131087845	0.286919831223629	0.566666666666667	0.903247631935047	0.192090395480226	0.980896399706098	0.192090395480226	0.566666666666667
GBM_rose	0.570871261378413	0.0692053198747058	0.136125654450262	0.866666666666667	0.558863328822733	0.0738636363636364	0.990407673860911	0.0738636363636364	0.866666666666667
NN_orig	0.960338101430429	-0.00128071036735158		0	0.999323410013532	0	0.960962914769031	0	0
NN_smote	0.823797139141743	0.145734184323902	0.200589970501475	0.566666666666667	0.834235453315291	0.121863799283154	0.97934868943606	0.121863799283154	0.566666666666667
NN_rose	0.747724317295189	0.106654610561866	0.167381974248927	0.65	0.751691474966171	0.0960591133004926	0.981448763250883	0.0960591133004926	0.65
RF_orig	0.960988296488947	0		0	1		0.960988296488947		0
RF_smote	0.888816644993498	0.220963642718769	0.266094420600858	0.516666666666667	0.903924221921516	0.179190751445087	0.978754578754579	0.179190751445087	0.516666666666667
RF_rose	0.592327698309493	0.0761019316738171	0.142270861833105	0.866666666666667	0.581190798376184	0.0774962742175857	0.990772779700115	0.0774962742175857	0.866666666666667

Showing 1 to 18 of 18 entries Previous 1 Next

Figure 13: Performance measures of the individual classifiers for the QPS case study, using 50% of training data and topics as input space

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Show 25 entries Search:

	Accuracy	Kappa	F1	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
svm_Linear_orig	0.960975609756098	0		0	1		0.960975609756098		0
svm_Linear_smote	0.821138211382114	0.194009578954893	0.246575342465753	0.75	0.824027072758037	0.147540983606557	0.987829614604463	0.147540983606557	0.75
svm_Linear_rose	0.738211382113821	0.131333070140808	0.190954773869347	0.791666666666667	0.736040609137056	0.108571428571429	0.988636363636364	0.108571428571429	0.791666666666667
svm_Poly_orig	0.960975609756098	0.0713476783691961	0.0769230769230769	0.0416666666666667	0.998307952622673	0.500000000000005	0.962479608482871	0.5	0.0416666666666667
svm_Poly_smote	0.895934959349593	0.22896097790315	0.272727272727273	0.5	0.912013536379019	0.1875	0.978221415607985	0.1875	0.5
svm_Poly_rose	0.744715447154472	0.144402796607916	0.203045685279188	0.833333333333333	0.741116751269036	0.115606936416185	0.990950226244344	0.115606936416185	0.833333333333333
svm_Radial_orig	0.960975609756098	0		0	1		0.960975609756098		0
svm_Radial_smote	0.856910569105691	0.204795909371419	0.254237288135593	0.625	0.866328257191201	0.159574468085106	0.982725527831094	0.159574468085106	0.625
svm_Radial_rose	0.960975609756098	0		0	1		0.960975609756098		0
GBM_orig	0.959349593495935	0.0659741206488062	0.0740740740740741	0.0416666666666667	0.996615905245347	0.333333333333333	0.962418300653595	0.333333333333333	0.0416666666666667
GBM_smote	0.876422764227642	0.159110535405873	0.208333333333333	0.416666666666667	0.895093062605753	0.138888888888889	0.974217311233886	0.138888888888889	0.416666666666667
GBM_rose	0.55609756097561	0.065989085265108	0.133333333333333	0.875	0.543147208121827	0.0721649484536083	0.990740740740741	0.0721649484536082	0.875
NN_orig	0.960975609756098	0		0	1		0.960975609756098		0
NN_smote	0.850406504065041	0.168650268888301	0.220338983050847	0.541666666666667	0.862944162436548	0.138297872340426	0.978886756238004	0.138297872340426	0.541666666666667
NN_rose	0.699186991869919	0.123870908124841	0.185022026431718	0.875	0.692047377326565	0.103448275862069	0.992718446601942	0.103448275862069	0.875
RF_orig	0.960975609756098	0		0	1		0.960975609756098		0
RF_smote	0.889430894308943	0.231955922865014	0.276595744680851	0.541666666666667	0.903553299492386	0.185714285714286	0.979816513761468	0.185714285714286	0.541666666666667
RF_rose	0.655284552845528	0.101880553833437	0.165354330708661	0.875	0.646362098138748	0.091304347826087	0.992207792207792	0.091304347826087	0.875

Showing 1 to 18 of 18 entries Previous 1 Next

Figure 14: Performance measures of the individual classifiers for the QPS case study, using 80% of training data and topics as input space

Show 25 entries Search:

	Accuracy	Kappa	F1	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
svm_Linear_orig	0.828978622327791	0		0	1		0.828978622327791		0
svm_Linear_smote	0.781472684085511	0.193755203996669	0.323529411764706	0.305555555555556	0.879656160458453	0.34375	0.859943977591036	0.34375	0.305555555555556
svm_Linear_rose	0.771971496437055	0.0961624474461045	0.225806451612903	0.194444444444444	0.891117478510029	0.269230769230769	0.842818428184282	0.269230769230769	0.194444444444444
svm_Poly_orig	0.826603325415677	-0.00470757461832732		0	0.997134670487106	0	0.828571428571429	0	0
svm_Poly_smote	0.767220902612827	0.196752589362199	0.337837837837838	0.347222222222222	0.853868194842407	0.328947368421053	0.863768115942029	0.328947368421053	0.347222222222222
svm_Poly_rose	0.748218527315915	0.0491349186056428	0.196969696969697	0.180555555555556	0.865329512893983	0.216666666666667	0.836565096952909	0.216666666666667	0.180555555555556
svm_Radial_orig	0.828978622327791	0		0	1		0.828978622327791		0
svm_Radial_smote	0.171021377672209	0	0.292089249492901	1	0	0.171021377672209		0.171021377672209	1
svm_Radial_rose	0.828978622327791	0		0	1		0.828978622327791		0
GBM_orig	0.809976247030879	0.0149166422930683	0.0697674418604651	0.0416666666666667	0.968481375358166	0.214285714285714	0.83046683046683	0.214285714285714	0.0416666666666667
GBM_smote	0.795724465558195	0.169739497339938	0.283333333333333	0.236111111111111	0.911174785100286	0.354166666666667	0.85254691689008	0.354166666666667	0.236111111111111
GBM_rose	0.171021377672209	0	0.292089249492901	1	0	0.171021377672209		0.171021377672209	1
NN_orig	0.795724465558195	0.190312192503802	0.306451612903226	0.263888888888889	0.905444126074499	0.365384615384615	0.856368563685637	0.365384615384615	0.263888888888889
NN_smote	0.700712589073634	0.135157167079692	0.315217391304348	0.402777777777778	0.762177650429799	0.258928571428571	0.86084142394822	0.258928571428571	0.402777777777778
NN_rose	0.605700712589074	0.0122678576476245	0.238532110091743	0.361111111111111	0.656160458452722	0.178082191780822	0.832727272727273	0.178082191780822	0.361111111111111
RF_orig	0.814726840855107	-0.00953025086079663	0.025	0.0138888888888889	0.979942693409742	0.125	0.828087167070218	0.125	0.0138888888888889
RF_smote	0.743467933491686	0.0851877188040722	0.23943661971831	0.236111111111111	0.848137535816619	0.242857142857143	0.843304843304843	0.242857142857143	0.236111111111111
RF_rose	0.171021377672209	0	0.292089249492901	1	0	0.171021377672209		0.171021377672209	1

Showing 1 to 18 of 18 entries Previous 1 Next

Figure 15: Performance measures of the individual classifiers for the ERIS case study, using 20% of training data and TDM as input space

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Show 25 entries Search:

	Accuracy	Kappa	F1	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
svm_Linear_orig	0.828897338403042	0		0	1		0.828897338403042		0
svm_Linear_smote	0.749049429657795	0.130622057497746	0.282608695652174	0.288888888888889	0.844036697247706	0.276595744680851	0.851851851851852	0.276595744680851	0.288888888888889
svm_Linear_rose	0.650190114068441	0.0639121015165582	0.26984126984127	0.377777777777778	0.706422018348624	0.209876543209877	0.846153846153846	0.209876543209877	0.377777777777778
svm_Poly_orig	0.828897338403042	0		0	1		0.828897338403042		0
svm_Poly_smote	0.775665399239544	0.106729606815958	0.233766233766234	0.2	0.894495412844037	0.28125	0.844155844155844	0.28125	0.2
svm_Poly_rose	0.703422053231939	0.083705556548151	0.264150943396226	0.311111111111111	0.784403669724771	0.229508196721311	0.846534653465347	0.229508196721311	0.311111111111111
svm_Radial_orig	0.828897338403042	0		0	1		0.828897338403042		0
svm_Radial_smote	0.171102661596958	0	0.292207792207792	1	0	0.171102661596958		0.171102661596958	1
svm_Radial_rose	0.828897338403042	0		0	1		0.828897338403042		0
GBM_orig	0.817490494296578	0.0573476702508961	0.111111111111111	0.0666666666666667	0.972477064220184	0.333333333333334	0.834645669291339	0.333333333333333	0.0666666666666667
GBM_smote	0.813688212927757	0.272619517977084	0.379746835443038	0.333333333333333	0.912844036697248	0.441176470588235	0.868995633187773	0.441176470588235	0.333333333333333
GBM_rose	0.216730038022814	0.00711065498662172	0.294520547945205	0.955555555555556	0.0642201834862385	0.174089068825911	0.875	0.174089068825911	0.955555555555556
NN_orig	0.783269961977186	0.119471365638767	0.24	0.2	0.903669724770642	0.3	0.84549356223176	0.3	0.2
NN_smote	0.749049429657795	0.173585983622167	0.326530612244898	0.355555555555556	0.830275229357798	0.30188679245283	0.861904761904762	0.30188679245283	0.355555555555556
NN_rose	0.726235741444867	0.167062549485353	0.333333333333333	0.4	0.793577981651376	0.285714285714286	0.865	0.285714285714286	0.4
RF_orig	0.828897338403042	0		0	1		0.828897338403042		0
RF_smote	0.817490494296578	0.165189789710356	0.25	0.177777777777778	0.94954128440367	0.421052631578947	0.848360655737705	0.421052631578947	0.177777777777778
RF_rose	0.171102661596958	0	0.292207792207792	1	0	0.171102661596958		0.171102661596958	1

Showing 1 to 18 of 18 entries Previous 1 Next

Figure 16: Performance measures of the individual classifiers for the ERIS case study, using 50% of training data and TDM as input space

Show 25 entries Search:

	Accuracy	Kappa	F1	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
svm_Linear_orig	0.828571428571429	0		0	1		0.828571428571429		0
svm_Linear_smote	0.638095238095238	0.0263543191800879	0.24	0.333333333333333	0.701149425287356	0.1875	0.835616438356164	0.1875	0.333333333333333
svm_Linear_rose	0.657142857142857	0.162234042553192	0.357142857142857	0.555555555555556	0.67816091954023	0.263157894736842	0.880597014925373	0.263157894736842	0.555555555555556
svm_Poly_orig	0.80952380952381	-0.0355029585798821		0	0.977011494252874	0	0.825242718446602	0	0
svm_Poly_smote	0.771428571428571	-0.0937500000000001		0	0.931034482758621	0	0.818181818181818	0	0
svm_Poly_rose	0.628571428571429	0.0172786177105831	0.235294117647059	0.333333333333333	0.689655172413793	0.181818181818182	0.833333333333333	0.181818181818182	0.333333333333333
svm_Radial_orig	0.828571428571429	0		0	1		0.828571428571429		0
svm_Radial_smote	0.171428571428571	0	0.292682926829268	1	0	0.171428571428571		0.171428571428571	1
svm_Radial_rose	0.828571428571429	0		0	1		0.828571428571429		0
GBM_orig	0.80952380952381	0.0885416666666665	0.166666666666667	0.111111111111111	0.954022988505747	0.333333333333333	0.838383838383838	0.333333333333333	0.111111111111111
GBM_smote	0.838095238095238	0.326160815402038	0.413793103448276	0.333333333333333	0.942528735632184	0.545454545454545	0.872340425531915	0.545454545454545	0.333333333333333
GBM_rose	0.180952380952381	0.00397088021178026	0.295081967213115	1	0.0114942528735632	0.173076923076923	1	0.173076923076923	1
NN_orig	0.79047619047619	0.150110375275938	0.266666666666667	0.222222222222222	0.908045977011494	0.333333333333333	0.849462365591398	0.333333333333333	0.222222222222222
NN_smote	0.752380952380952	0.258957654723127	0.409090909090909	0.5	0.804597701149425	0.346153846153846	0.886075949367089	0.346153846153846	0.5
NN_rose	0.676190476190476	0.128843338213763	0.32	0.444444444444444	0.724137931034483	0.25	0.863013698630137	0.25	0.444444444444444
RF_orig	0.79047619047619	-0.066481994459834		0	0.954022988505747	0	0.821782178217822	0	0
RF_smote	0.847619047619048	0.312039312039312	0.384615384615385	0.277777777777778	0.96551724137931	0.625	0.865979381443299	0.625	0.277777777777778
RF_rose	0.180952380952381	0.00397088021178026	0.295081967213115	1	0.0114942528735632	0.173076923076923	1	0.173076923076923	1

Showing 1 to 18 of 18 entries Previous 1 Next

Figure 17: Performance measures of the individual classifiers for the ERIS case study, using 80% of training data and TDM as input space

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Show 25 entries Search:

	Accuracy	Kappa	F1	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
svm_Linear_orig	0.828978622327791	0		0	1		0.828978622327791		0
svm_Linear_smote	0.65083135391924	0.114926419060967	0.316279069767442	0.472222222222222	0.687679083094556	0.237762237762238	0.863309352517986	0.237762237762238	0.472222222222222
svm_Linear_rose	0.534441805225653	0.0674261431703624	0.304964539007092	0.597222222222222	0.521489971346705	0.204761904761905	0.862559241706161	0.204761904761905	0.597222222222222
svm_Poly_orig	0.828978622327791	0		0	1		0.828978622327791		0
svm_Poly_smote	0.752969121140142	0.191192226696715	0.341772151898734	0.375	0.830945558739255	0.313953488372093	0.865671641791045	0.313953488372093	0.375
svm_Poly_rose	0.648456057007126	0.146114841715774	0.345132743362832	0.541666666666667	0.670487106017192	0.253246753246753	0.876404494382022	0.253246753246753	0.541666666666667
svm_Radial_orig	0.828978622327791	0		0	1		0.828978622327791		0
svm_Radial_smote	0.171021377672209	0	0.292089249492901	1	0	0.171021377672209		0.171021377672209	1
svm_Radial_rose	0.828978622327791	0		0	1		0.828978622327791		0
GBM_orig	0.817102137767221	0.142203170066947	0.222222222222222	0.152777777777778	0.954154727793696	0.407407407407407	0.845177664974619	0.407407407407407	0.152777777777778
GBM_smote	0.68646080760095	0.10208434318953	0.290322580645161	0.375	0.750716332378224	0.236842105263158	0.853420195439739	0.236842105263158	0.375
GBM_rose	0.486935866983373	0.0440870387890256	0.294117647058824	0.625	0.458452722063037	0.192307692307692	0.855614973262032	0.192307692307692	0.625
NN_orig	0.828978622327791	0		0	1		0.828978622327791		0
NN_smote	0.828978622327791	0		0	1		0.828978622327791		0
NN_rose	0.508313539192399	0.0478755367151395	0.293515358361775	0.597222222222222	0.489971346704871	0.194570135746606	0.855	0.194570135746606	0.597222222222222
RF_orig	0.831353919239905	0.0731760255495957	0.101265822784810	0.0555555555555556	0.991404011461318	0.571428571428572	0.835748792270531	0.571428571428571	0.0555555555555556
RF_smote	0.793349168646081	0.109157241882524	0.216216216216216	0.166666666666667	0.922636103151863	0.307692307692308	0.842931937172775	0.307692307692308	0.166666666666667
RF_rose	0.527315914489311	0.0790986534762298	0.316151202749141	0.638888888888889	0.504297994269341	0.210045662100457	0.871287128712871	0.210045662100457	0.638888888888889

Showing 1 to 18 of 18 entries Previous 1 Next

Figure 18: Performance measures of the individual classifiers for the ERIS case study, using 20% of training data and topics as input space

Show 25 entries Search:

	Accuracy	Kappa	F1	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
svm_Linear_orig	0.828897338403042	0		0	1		0.828897338403042		0
svm_Linear_smote	0.695817490494297	0.28210727446431	0.452054794520548	0.733333333333333	0.688073394495413	0.326732673267327	0.925925925925926	0.326732673267327	0.733333333333333
svm_Linear_rose	0.669201520912547	0.205217270485254	0.391608391608392	0.622222222222222	0.678899082568807	0.285714285714286	0.896969696969697	0.285714285714286	0.622222222222222
svm_Poly_orig	0.828897338403042	0		0	1		0.828897338403042		0
svm_Poly_smote	0.722433460076046	0.107977512428565	0.277227722772277	0.311111111111111	0.807339449541284	0.25	0.85024154589372	0.25	0.311111111111111
svm_Poly_rose	0.574144486692015	0.119665271966527	0.341176470588235	0.644444444444444	0.559633027522936	0.232	0.884057971014493	0.232	0.644444444444444
svm_Radial_orig	0.828897338403042	0		0	1		0.828897338403042		0
svm_Radial_smote	0.171102661596958	0	0.292207792207792	1	0	0.171102661596958		0.171102661596958	1
svm_Radial_rose	0.828897338403042	0		0	1		0.828897338403042		0
GBM_orig	0.798479087452472	-0.00201279562935805	0.0701754385964912	0.0444444444444444	0.954128440366973	0.166666666666667	0.828685258964143	0.166666666666667	0.0444444444444444
GBM_smote	0.730038022813688	0.103595602707503	0.268041237113402	0.288888888888889	0.821100917431193	0.25	0.848341232227488	0.25	0.288888888888889
GBM_rose	0.623574144486692	0.156423132998542	0.361290322580645	0.622222222222222	0.623853211009174	0.254545454545455	0.888888888888889	0.254545454545455	0.622222222222222
NN_orig	0.828897338403042	0		0	1		0.828897338403042		0
NN_smote	0.752851711026616	0.242276494836222	0.392523364485981	0.466666666666667	0.811926605504587	0.338709677419355	0.880597014925373	0.338709677419355	0.466666666666667
NN_rose	0.737642585551331	0.293203505355404	0.448	0.622222222222222	0.761467889908257	0.35	0.907103825136612	0.35	0.622222222222222
RF_orig	0.828897338403042	0		0	1		0.828897338403042		0
RF_smote	0.76425855513308	0.0323997151673392	0.162162162162162	0.133333333333333	0.894495412844037	0.206896551724138	0.833333333333333	0.206896551724138	0.133333333333333
RF_rose	0.475285171102662	0.0692891578623449	0.316831683168317	0.711111111111111	0.426605504587156	0.203821656050955	0.877358490566038	0.203821656050955	0.711111111111111

Showing 1 to 18 of 18 entries Previous 1 Next

Figure 19: Performance measures of the individual classifiers for the ERIS case study, using 50% of training data and topics as input space

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.

Show 25 entries Search:

	Accuracy	Kappa	F1	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall
svm_Linear_orig	0.828571428571429	0		0	1		0.828571428571429		0
svm_Linear_smote	0.685714285714286	0.109483423284503	0.297872340425532	0.388888888888889	0.747126436781609	0.241379310344828	0.855263157894737	0.241379310344828	0.388888888888889
svm_Linear_rose	0.60952380952381	0.161309175920514	0.369230769230769	0.666666666666667	0.597701149425287	0.25531914893617	0.896551724137931	0.25531914893617	0.666666666666667
svm_Poly_orig	0.828571428571429	0		0	1		0.828571428571429		0
svm_Poly_smote	0.704761904761905	0.163454124903624	0.340425531914894	0.444444444444444	0.758620689655172	0.275862068965517	0.868421052631579	0.275862068965517	0.444444444444444
svm_Poly_rose	0.6	0.152249134948097	0.363636363636364	0.666666666666667	0.586206896551724	0.25	0.894736842105263	0.25	0.666666666666667
svm_Radial_orig	0.828571428571429	0		0	1		0.828571428571429		0
svm_Radial_smote	0.171428571428571	0	0.292682926829268	1	0	0.171428571428571		0.171428571428571	1
svm_Radial_rose	0.828571428571429	0		0	1		0.828571428571429		0
GBM_orig	0.80952380952381	-0.0355029585798821		0	0.977011494252874	0	0.825242718446602	0	0
GBM_smote	0.714285714285714	-0.00574712643678207	0.166666666666667	0.166666666666667	0.827586206896552	0.166666666666667	0.827586206896552	0.166666666666667	0.166666666666667
GBM_rose	0.542857142857143	0.102564102564102	0.333333333333333	0.666666666666667	0.517241379310345	0.222222222222222	0.882352941176471	0.222222222222222	0.666666666666667
NN_orig	0.828571428571429	0		0	1		0.828571428571429		0
NN_smote	0.752380952380952	0.258957654723127	0.409090909090909	0.5	0.804597701149425	0.346153846153846	0.886075949367089	0.346153846153846	0.5
NN_rose	0.685714285714286	0.168466522678186	0.352941176470588	0.5	0.724137931034483	0.272727272727273	0.875	0.272727272727273	0.5
RF_orig	0.828571428571429	0		0	1		0.828571428571429		0
RF_smote	0.761904761904762	0.179943767572633	0.324324324324324	0.333333333333333	0.850574712643678	0.31578947368421	0.86046511627907	0.315789473684211	0.333333333333333
RF_rose	0.647619047619048	0.222222222222222	0.412698412698413	0.722222222222222	0.632183908045977	0.288888888888889	0.916666666666667	0.288888888888889	0.722222222222222

Showing 1 to 18 of 18 entries Previous 1 Next

Figure 20: Performance measures of the individual classifiers for the ERIS case study, using 80% of training data and topics as input space

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the author.