Bug Report Severity Level Prediction in Free Libre/Open Source Software: A Survey and New Perspectives

Luiz Alberto Ferreira Gomes
Institute of Computing (IC)
University of Campinas (UNICAMP)
Campinas, Brazil
gomes.luiz@ic.unicamp.br

Ricardo da Silva Torres
Institute of Computing (IC)
University of Campinas (UNICAMP)
Campinas, Brazil
rtorres@ic.unicamp.br

Mario Lúcio Côrtes
Institute of Computing (IC)
University of Campinas (UNICAMP)
Campinas, Brazil
cortes@ic.unicamp.br

Context: The severity level attribute of a bug report is considered one of the most critical variables for planning evolution and maintenance in Free/Libre Open Source Software. This variable measures the impact the bug has on the successful execution of the software system and how soon a bug needs to be addressed by the development team. Both business and academic community have made an extensive investigation towards the proposal to provide methods to automate the bug report severity prediction.

Objective: This paper aims to provide a comprehensive review of recent research efforts on automatically bug report severity prediction. To the best of our knowledge, this is the first review to categorize quantitatively more than ten aspects of the experiments reported in several papers on bug report severity prediction.

Method: The mapping review was performed by searching four electronic databases. It was considered studies published until December 2017. The initial resulting set was comprised of 54 papers. From this set, a total of 18 papers were selected. After performing snowballing more nine papers were selected.

Results: From the mapping study, we identified 27 studies addressing bug report severity prediction on Free/Libre Open Source Software.

Conclusion: The gathered data confirm the relevance of this topic, reflects the scientific maturity of the research area, as well as, identify gaps, which can motivate new research initiatives. Furthermore, the results presented in this review demonstrate that most of the surveyed papers proposed methods to predict severity level which extract features from unstructured text information. At the same time, these results showed that traditional machine learning algorithms and text mining methods have been playing a central role in performed experiments. The review suggests that there is room for improving prediction results using state-of-the-art machine learning and text mining algorithms and techniques.

Index Terms

Software maintenance, bug tracking systems, bug reports, severity level prediction, software repositories, systematic mapping, machine learning.

I. Introduction

Bug Tracking Systems (BTS) have been playing a key role as a communication and collaboration tool in Closed Source Software (CSS) and Free/Libre Open Source Software (FLOSS). In both development environments, planning of software evolution and maintenance activities relies primarily on information of bug reports registered in this kind of system. This is particularly true in FLOSS, which is characterized by the existence of many users and developers with different levels of expertise spread out around the world, who might create or be responsible for dealing with several bug reports [**?**].

A user interacts with a BTS often through a simple mechanism called bug report form, which enables to communicate a bug to those in charge of developing or maintaining the software system [**?**]. Initially, he or she should inform a short description, a long description, and an associated severity level (e.g., blocker, critical, major, minor, and trivial). Subsequently, a software team member reviews this bug report and confirms or declines it (e.g., due to bug report duplication). If the bug report is confirmed, the team member should provide more information to complement the bug report form, for example, by indicating its priority and by assigning a person who will be responsible for fixing the bug.

The information of severity level is recognized as a critical variable for prioritizing and planing how bug reports will be dealt with [**?**]. It measures the impact the bug has on the successful execution of the software system and defines how soon the bug needs to be addressed [**?**]. However, the severity level assignment remains mostly a manual process, which relies only on the experience and expertise of the person who has opened the bug report [**?**], [**?**], [**?**].

Moreover, the number of bug reports in large and medium software FLOSS projects is frequently very large [**?**]. For example, Eclipse project had 84,245 bug reports opened from 2013 to 2015 alone, whereas Android project had over 107,456, and JBoss

project had over 81,920. Therefore, a manual assignment of severity level may be a quite subjective, cumbersome and error-prone process, and a wrong decision throughout bug report lifecycle may strongly disturb the planning of maintenance activities. For instance, an important maintenance team resource could be allocated to address less significant bug reports before the most important ones.

Due to the evident importance of bug report severity information for the planning of FLOSS maintenance, both business and academic community have demonstrated great interest, and plenty of research has been done in this field. However, there are few mapping reviews well characterizing this area, positioning existing works and measuring current progress. Although these existing works helped to understand the challenges and opportunities in this research area [**?**], [**?**], they suffer from one shortcoming: existing reviews lack in-depth coverage of the different dimensions of bug report severity prediction. Our current systematic mapping review fills this gap, by investigating and analyzing relevant papers — published from 2010 to 2017 — related to bug report severity prediction. To the best of our knowledge, this is the first review to provide a detailed document with the analysis of more than ten relevant aspects of the surveyed experiments, including: granularity of output class, FLOSS repositories, features and feature selection methods, text mining methods, machine learning algorithms, performance evaluation measures, sampling techniques, statistical tests, experimental tools, and type of solution.

A systematic mapping review aims to characterize the state-of-the-art on bug report severity prediction. This sort of study provides a broad overview of a research area to determine whether there is research evidence on a particular topic. According to Kitchenham et al. [**?**] and Petersen et al. [**?**], results yielded by a mapping review help recognizing gaps to suggest future research and provide a direction to engage in new research activities appropriately. The key contributions of our review are three-fold:

- It proposes a categorization scheme to organize the experiments reported in papers on bug report severity prediction. For example, it provides a taxonomy to categorize the severity prediction problem type based on the granularity severity level which one wishes to predict.
- It provides an overview of the state-of-the-art research by summarizing patterns, procedures, and methods employed by researches to predict bug report severity. For example, it indicates which are the most used machine learning algorithms in reported experiments.
- It discusses challenges, and open issues and future directions in this research field to share the vision and expand the horizon of bug report severity prediction.

The remaining of this paper is organized as follows: Section II provides the basic information background necessary to understand the research area. Section III presents related work. Section IV describes our research method. Section V presents our results. Section VI presents final findings and discussion. Finally, Section VII concludes the paper.

II. Terminology and Background

This section presents an overview on general concepts necessary to understand this research area, namely Bug Tracking Systems, Machine Learning (algorithms, feature selection methods, evaluation measures, and sampling methods), and Statistical Tests. More specific concepts will be detailed as they are cited in the document.

A. Bug Tracking System (BTS)

BTS [**?**] is a software application that keeps the record and tracks information about change requests, bug fixes, and technical support that could occur during the software lifecycle. Usually, while reporting a bug in a BTS, a user is asked to provide information about the bug by filling out a form, typically called bug report form.

1) Bug Report: Although there is no agreement on the terminology or the amount of information that users must provide to fill a typical bug report (illustrated in Figure 1, the example refers to a bug report extracted from Bugzilla of Eclipse project.), they often describe their needs in popular BTS (e.g., Bugzilla, Jira, and Redmine) [**?**] providing information about at least attributes shown in Table I.

TABLE I: Commom attributes in a bug report form.

Type	Type of report (e.g., bug, improvement, and new feature)
Summary	Short description of report in one line.
Description	Long and detailed description of report in many lines of text. It could include source code snippets and stack tracing reports.
Severity	Level of severity of report (e.g., blocker, critical, major, minor and trivial).

After the user has reported a bug, the development team is in charge of its assessment. The assessment consists of approving or, for some reason (e.g., duplication), not approving the bug. In case of approval, this team may provide complementary information by, for example, assigning a person to be responsible for handling this request or defining a new severity level

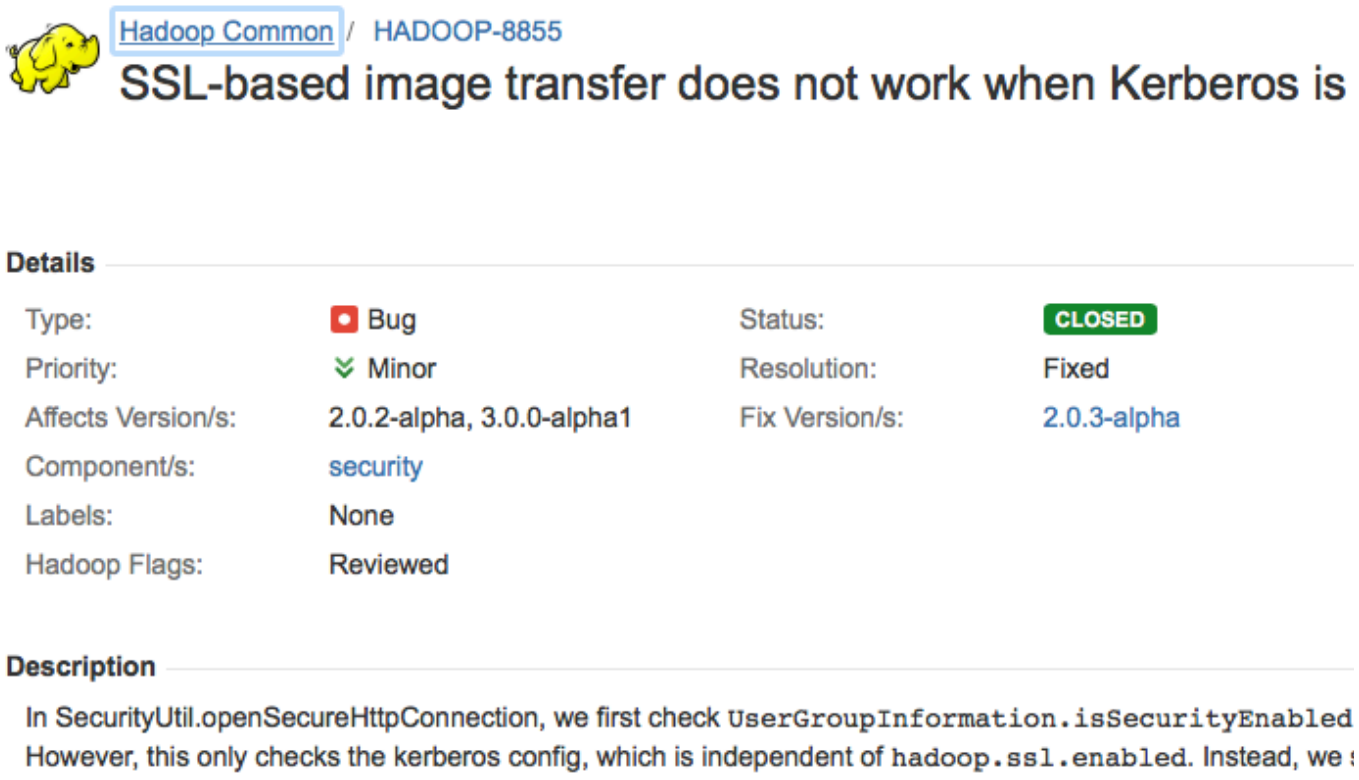

Fig. 1: A typical bug report (https://issues.apache.org/jira/browse/HADOOP-8855, as of September 2018)

for the report. Typically, the sequence of steps a bug report goes through is modeled as a state machine. Figure 2a shows an example of such a state machine, with a typical set of states a bug report can hold during its lifecycle in a BTS. Initially, a bug report is in *Unconfirmed* state. The developer team will change the bug report state to *Resolved*, if the bug were not confirmed, or, otherwise, to *New*. When someone was in charge to fixing the bug, the bug report state will be changed to *Assigned* by the developer team. Therefore, in the standard flow, the bug report status will be assigned to resolved (bug fixed), then verified (bug checked), and finally closed. As shown in Figure 2a, others state transitions may occur throughout the bug report lifecycle. All changes occurred in a bug report are stored in a repository, keeping valuable historical information about a particular software.

Both users and development team members can define or redefine the severity level for a bug during the lifecycle. Figure 2b illustrates Bugzilla guidelines for assigning bug severity level. The Figure shows that such choices should be based on an affirmative answer to a question which characterizes a severity level appropriately. Also, the Figure indicates that *Trivial* and *Blocker* are lower and higher respectively.

To predict severity level, researchers sometimes aggregate these levels in severe (blocker, critical, and major) and non-severe (normal, minor and trivial) to work with a coarse-grained classification problem. Furthermore, some of them ignore the default severity level (often "normal") because they consider this level as a choice made by users when they are not sure about the correct severity level. Other studies choose to predict a bug report severity level as *blocking* or *non-blocking* bug. A blocking bug is one that prevents other bugs to be fixed [**?**].

B. Machine Learning

Machine Learning (ML) [**?**] is an application of Artificial Intelligence (AI) that provides systems the ability to learn and improve from experience without being explicitly programmed. There are two types of ML algorithms: predictive (or supervised) and descriptive (or unsupervised). A predictive algorithm builds a model based on historical training data and uses this model to predict, from the values of input attributes, an output label (class attribute) for a new sample. A predictive task is called classification when the label value is discrete, or regression when the label value is continuous.

On the other hand, a descriptive algorithm explores or describes a dataset. There is not an output label associated with a sample. Data clustering and pattern discovery are two examples of descriptive tasks. Bug report severity prediction is considered a classification problem; therefore, more detailing of descriptive algorithms are outside the scope of this paper.

1) ML algorithms: A ML algorithm works over a dataset, which contains many samples or instances x_i, where $i = \{1..n\}$. Each instance is composed of $\{x_{i1}, x_{i2}, ..., x_{id}\}$ input attributes or independent variables, where $d = \{1..m\}$, and one output

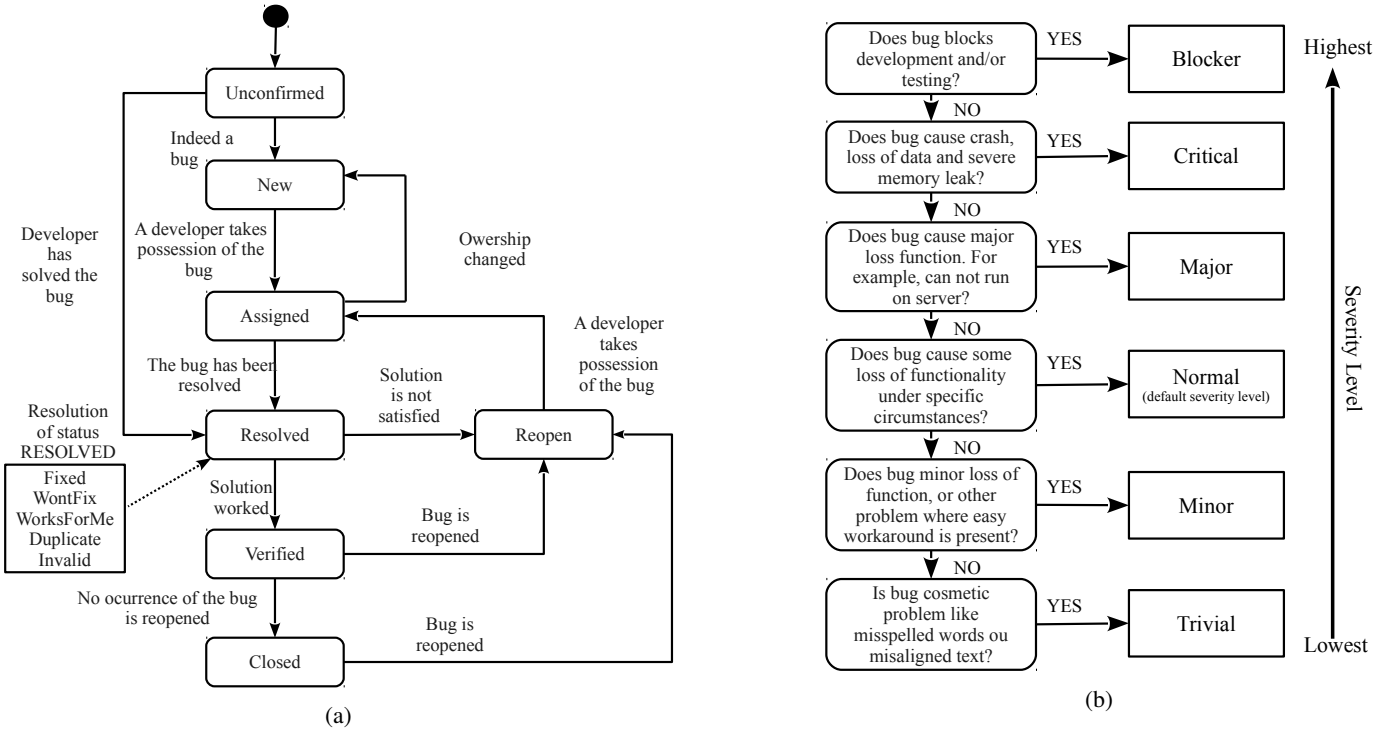

Fig. 2: (a) The bug report lifecycle according to Zhang et al. [**?**]. (b) Bugzilla guideline for bug report severity level assignment (https://www.bugzilla.org, as of September 10, 2018).

attribute or dependent variable, $x_{i(m+1)}$. Input attributes are commonly named features or feature vector, and output attribute as commonly named class or category. The most traditional ML classification algorithms are k-Nearest Neighbors, Näive Bayes, Decision Tree, Neural Networks, Random Forest and Support Vector Machine. In practice, they can be applied for both classification and regression tasks. However, this mapping review regards them only in the classification scenario. Next, a brief description of each one is presented [**?**]:

- **k-Nearest Neighbors** (k-NN) process the available instances (or neighbors) in a dataset and classifies a new instance based on its similarity measure to the k-nearest neighbors. Usually, k-NN algorithms utilize a Euclidean distance to quantify the proximity of neighbors. To calculate this distance, each feature vector of each instance in a dataset should represent a point of an n-dimensional space.
- **Näive Bayes** (NB) decides to which class an instance belongs based on the Bayesian Theorem of conditional probability. The probabilities of an instance belonging to each of the C_k classes given the instance x is $P(C_k|x)$. Näive Bayes classifiers assume that given the class variable, the value of a particular feature is independent of the value of any other feature.
- **Decision Tree** consists of a collection of internal nodes and leaf nodes in a tree, organized in a hierarchical model. Each internal node represents a feature, and each leaf node corresponds to a class label. The decision tree classifiers organize a series of test questions and conditions in a tree structure. A decision tree represents the model capable of guiding the decision making on the determination of the class to which an individual belongs.
- **Neural Network** is a learning algorithm that is inspired by the structure and functional aspects of biological neural networks [**?**]. It structured as a network of units called neurons, with weighted, directed connections. Neural network models have been demonstrated to be capable of achieving remarkable performance in document classification [**?**].
- **Support Vector Machine(SVM)**Each feature vector of each instance is a point in an n-dimensional space. SVM learns in this space an optimal way to separate the training instances according to their class labels. The output of this algorithm is a hyperplane, which maximizes the separation among feature vectors of instances of different classes. Given a new instance, SVM assigns a label based on which subspace its feature vector belongs to [**?**].
- **Random Forest** [**?**] relies on two core principles: (i) the creation of hundreds of decision trees and their combination into a single model; and (ii) the final decision is based on the ruling of the majority of the forming trees.

2) Feature Selection Methods: Feature selection is the process of choosing a subset of features that better contribute to the accuracy of a predictive model. Three typical feature selection methods are described next [**?**]:

- **Information Gain (IG)**: this method measures the number of bits of information obtained for category prediction by knowing the presence or absence of a feature in a dataset.

- **Chi-square (CHI)**: this method measures the lack of independence between a feature f and category c_i and can be compared to the chi-square distribution with one degree of freedom.
- **Correlation Coefficient (CC)**: this method defines the correlation coefficient of feature f with a category c_i.

3) Evaluation measures: Accuracy, precision, recall, and F-measure are four measures commonly used to evaluate the performance of prediction models [**?**]. The computation of the values of these measures are based on a *confusion matrix* [**?**], which represents the number of true/false positives, and the number of true/false negatives for each instance class value when making a prediction. Each measure is described next [**?**]:

- **Accuracy** is the percentage of correctly classified observations among all observations:

$$Accuracy = \frac{TP + TN}{P + N} \tag{1}$$

  where P is the total of positive class instances, N is the total of negative class instances, TP is the number of true positives, and TN is the number of true negatives.
- **Precision** is the percentage of correctly classified observations among all observations that were assigned to the class by the classifier. It can be viewed as a measure of classifier exactness. A low precision can also indicate a large number of false positives. More formally recall is defined as:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

  where TP is the number of true positives, and FP is the number of false positives.
- **Recall** of a classification method can be defined as the percentage of correctly classified observations among all observations belonging to that class. It can be viewed as a measure of a classifiers completeness. A low recall indicates many false negatives in testing classification step. More formally recall is defined as:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

  where TP is the number of true positives, and FN is the number of false negatives.
- **F-measure** is a harmonic mean of the precision and recall [**?**]. F-measure can be calculated as:

$$\textit{F-measure} = \frac{2 \times (precision \times recall)}{(precision + recall)} \tag{4}$$

Receiving Operating Characteristics (ROC) is an alternative measure to evaluate binary classifiers. A ROC curve [**?**] is a bidimensional chart, where the X-axis represents false positives, and the Y-axis represents true positives. The Area Under ROC Curve (AUC), ranging between 0 and 1, is used to assess the performance of ML algorithms. An algorithm outperforms another one if its AUC value is closer to 1.

4) Sampling methods: The evaluation of the supervised method effectiveness is mainly based on two datasets with labeled samples, one for training the predictive model and the other for testing this model. Assessing performance of a predictive model using the same dataset for training and testing is not recommend and may yield misleading optimistic estimations [**?**]. In order to obtain more reliable predictive estimates, resampling methods can be used to split the entire dataset into a training dataset and a testing dataset. Such methods according to Facelli et al. [**?**] include:

- **Holdout** splits a dataset into a ratio of *p* for training and $(1 - p)$ for testing.
- **Cross-Validation (CV)** divides a dataset into k folds. In each iteration, it saves a different fold for testing and uses all the others for training.
- **Bootstrap** generates *r* training subsets from the original dataset. It randomly samples instances with replacement from this set. Unselected cases make up the test subset. The result is the average performance observed for each test subset.

5) Statistical tests: Many scenarios require running several ML algorithms to chose the best predictive model. Even though the performance of these algorithms may be shown to be different on specific datasets, it needs to be confirmed whether the observed differences are statistically significant and not merely coincidental [**?**]. In this situation, conducting statistical tests is a recommended practice for reliable comparison between predictive models under investigation [**?**]. A brief description of four common statistical tests is provided:

- **T-Test** [**?**] is a parametric statistical hypothesis test that can be used to assess whether the means of two groups are statistically different from each other.
- **Wilcoxon signed-rank test** [**?**] is a non-parametric statistical hypothesis test that can be used to determine whether two dependent samples, selected from the population, have the same distribution.
- **Proportion test** [**?**] is a parametric statistical hypothesis test that can be used to assess whether or not a sample from a population represents the true proportion of the entire population.
- **Shapiro Wilk test** [**?**] is a parametric statistical hypothesis test that can be used to test whether a sample $x_1, ..., x_n$ came from a normal distribution of the population.

C. Text Mining

The common ML algorithms cannot directly process unstructured text (e.g., *summary* and *description* fields from bug report form). Therefore, during a preprocessing step, these unstructured text fields are converted into more manageable representations. Typically, the content of these fields is represented by feature vectors, points of an n-dimensional space. Text mining is the process of converting unstructured text into a structure suited to analysis [**?**]. It is composed of three primary activities [**?**]:

- **Tokenization** is the action to parsing a character stream into a sequence of tokens by splitting the stream at delimiters. A token is a block of text or a string of characters (without delimiters such as spaces and punctuation), which is a useful portion of the unstructured data.
- **Stop words removal** eliminates commonly used words that do not provide relevant information to a particular context, including prepositions, conjunctions, articles, common verbs, nouns, pronouns, adverbs, and adjectives.
- **Stemming** is the process of reducing or normalizing inflected (or sometimes derived) words to their word stem, base form—generally a written word form (e.g., "working" and "worked" into "work").

Two of the most traditional ways of representing a document relies on the use of a bag of words (unigrams) or a bag of bigrams (when two terms appear consecutively, one after the other) [**?**]. In this approach all terms represent features, and thus the dimension of the feature space is equal to the number of different terms in all documents (bug reports). Methods for assigning weights to features may vary. The simplest one is to assign binary values representing the presence or absence of the term in each text. Term Frequency (TF), another type of quantification scheme, considers the number of times in which the term appears in each document. Term Frequency-Inverse Document Frequency (TD-IDF), a more complex type of scheme, takes into account the frequencies of the term in each document, and in the whole collection. The importance of a term in this scheme is proportional to its frequency in the document and inversely proportional to the frequency of the term in the collection.

III. Related Work

To the extent of our knowledge, only two papers [**?**], [**?**] have reviewed the literature about bug report severity prediction. Cavalcanti et al. [**?**] performed a review of 142 papers, published between 2000 to 2012, that investigated challenges and opportunities for software change repositories. Just seven of them are related to change request prioritization, which is defined in Bugzilla by two fields: *priority* and *severity*. Four out of them addressed bug report severity prediction. Only two papers (Lamkanfi et al. [**?**] and Lamkanfi et al. [**?**]), however, addressed severity prediction on FLOSS projects. That mapping review shows that all of seven papers used some Information Retrieval Model: one paper used the vector space representation (binary and term count), and all seven used TF-IDF. It also shows that all papers implement learning techniques such as SVM (4 out of 7), Decision tree (2 out of 7), k-NN (2 out of 7), Naïve Bayes (3 out of 7), and Naïve Bayes Multinomial (1 out of 7).

The review presented by Uddin et al. [**?**], in turn, surveyed published papers in bug prioritization. The authors reviewed and analyzed in depth 32 distinct papers published between 2003 and 2015. The aim of that analysis was "to summarize the existing work on bug prioritization and some problems in working with bug prioritization". That work categorizes research initiatives according to ML algorithms, evaluation measures, data sets, researchers, and publication venue. It can be worth to note that the authors investigated only eight papers about predicting of severity level. In contrast, the current mapping review investigated in depth 27 papers about bug report severity prediction published from 2010 to 2017.

Although these reviews present relevant results for research in this area, they have one difference to our work. There is no explicit focus on bug report severity prediction. In fact, those papers perform a brief review of this topic by addressing mainly bug report priority prediction. The current review has a broader goal of mapping studies addressing many research questions and concepts related to bug report severity prediction. Furthermore, this review considers relevant aspects (e.g., sampling techniques and statistical tests), not addressed before, in the characterization of papers.

IV. Research method

This section describes the research method used in this mapping review for identifying and analyzing relevant papers. It is mainly based on the software engineering systematic literature review guidelines and recommendations proposed by Kitchenham et al. [**?**], Brhel et al. [**?**], Souza et al. [**?**], and Petersen et al. [**?**]. These authors suggest a process with three main steps: planning, conducting, and reporting. In the planning step, they recommend defining a review protocol that includes a set of research questions, inclusion and exclusion criteria, sources of papers, search string, and mapping procedures. In the conducting step, they recommend selecting and retrieving papers for data extraction. Finally, in the reporting step, they recommend analyzing the results used to address research questions defined before.

A. Research questions

The main goal of this mapping review is to provide a current status of the research on bug report severity prediction in FLOSS. To ensure an unbiased selection process, this review addresses the following research questions.

- **RQ$_1$. When and where have the studies been published?** This research question gives to the researcher a perception whether the topic of this mapping review seems to be broad and new, providing an overview about where and when papers were published.
- **RQ$_2$. What FLOSS are the most used as report sources in experiments for bug report severity prediction?** This research question names the FLOSS used as report source in problems of bug report severity prediction. This overview can be useful for researchers that intend to accomplish new initiatives in this area, and can also motivate the adoption of new FLOSS in future research to bridge existing gaps.
- **RQ$_3$. Was bug report severity prediction most addressed as either a fine-grained label or coarse-grained label prediction problem?** This research question investigates whether bug report severity prediction was handled as a fine-grained label (i.e., multi-label) or as a coarse-grained label (i.e., two-label or three-label) prediction problem. Understanding each solution provided for each prediction problem type is essential for guiding the selection of suitable machine learning algorithm.
- **RQ$_4$. What are the most common features used for bug report severity prediction?** This question investigates features used for bug report severity prediction in FLOSS. It can help to map which features have been considered more effective for this prediction problem.
- **RQ$_5$. What are the most common feature selection methods used for bug report severity prediction?** This question complements the previous one, looking for feature selection methods employed in the papers for bug report severity prediction in FLOSS. Also, this RQ allows for identifying which feature selection approaches have been considered more suitable for this prediction problem.
- **RQ$_6$. What are the most used text mining or information retrieval methods for bug report severity prediction?** This research question indicates the main text mining approaches used to extract features from textual fields of bug reports. It can guide researchers in the task of selecting text mining methods more suitable to be applied in bug report severity prediction or similar problems.
- **RQ$_7$. What are the most used machine learning algorithms for bug report severity prediction?** This research question aims to identify the main ML algorithms, which have been adopted for bug report severity prediction in FLOSS. Also, this RQ can be useful for any researcher that intends to accomplish further initiatives in this area, as well as to guide him/her in search of other algorithms to advance the state of the art.
- **RQ$_8$. What are the measures typically used to evaluate ML algorithms performance for bug report severity prediction?** This research question aims to find out which measures are used to evaluate ML algorithms performance. It can be helpful to identify the more appropriate measures to evaluate ML algorithms performance for bug report severity prediction.
- **RQ$_9$. Which sampling techniques are applied most frequently to generate more reliable predictive performance estimates in severity prediction of a bug report?** This question complements the previous one, investigating sampling techniques, which have been used to generate more reliable and accurate ML models. It can be useful to analyze sampling techniques, which might be more suitable to improve machine learning algorithms performance for bug report severity prediction.
- **RQ$_{10}$. Which statistical tests were used to compare the performance between two or more ML algorithms for bug report severity prediction?** The answer to this research question can be useful to identify which statistical tests were used in the context of bug report severity prediction. In addition, it allows grasping existing protocols for the comparison between two or more ML algorithms performance using statistical significance tests.
- **RQ$_{11}$. Which software tools were used to run experiments for bug report severity prediction?** This research question highlights the software tools most used to run experiments for bug reports severity prediction in FLOSS projects. It is useful to provide information for researchers and practitioners about technologies that could be used in their experiments.
- **RQ$_{12}$. Which solution types were proposed for the problem of bug report severity prediction?** This research question examines whether the most common solutions proposed in the selected studies are either online or off-line. This is important to separate the solutions that can be applied only in an experimental environment (off-line) from those that also can be deployed in a real scenario (online).

B. Paper selection

This section describes the selection process carried out through this systematic mapping review. It comprises four steps: (i) terms and searching string; (ii) sources for searching; (iii) inclusion and exclusion criteria; and (iv) data storage procedures.

1) Terms and search string: The base string was constructed from three main search terms related to three distinct knowledge areas: “open source project”, “bug report” and “severity predict.” To build the search string, terms were combined with “AND” connectors. The search string syntax was adapted according to particularities of each source (e.g., wildcards, connectors, apostrophes, and quotation marks) before it has been applied on three metadata papers: title, abstract, and keywords. Table II exhibits terms and final search string used in this systematic mapping review.

TABLE II: Search string.

Area	Search term
Bug report	"bug report"
Open source	"open source software"
Severity prediction	"severity predict"
Search string:	"open source software" AND "severity predict" AND "bug report"

2) Sources: To accomplish this systematic mapping review, four electronic databases and one popular search engines were selected, as recommended by Kitchenham et al. [**?**]. Table III shows each selected search sources, as well as the search date, and the period covered by the search.

TABLE III: Search sources.

Source	Electronic address	Type	Search Date	Years covered
ACM	http://dl.acm.og	Digital library	Dec,17	2010 - 2017
IEEE	http://ieeexplore.ieee.org	Digital library	Dec,17	2010 - 2017
Google Scholar	http://www.scholar.google.com	Search engine	Dec,17	2010 - 2017
Science Direct	http://www.sciencedirect.com	Digital library	Dec,17	2010 - 2017
SpringLink	http://www.springerlink.com	Digital library	Dec,17	2010 - 2017

3) Inclusion and exclusion criteria: The inclusion criteria below allow the identification of papers in the existing literature.

IC$_1$. The paper discusses bug report severity prediction in FLOSS projects.

On the contrary, the following exclusion criteria allow excluding all papers that satisfy any of them.

EC$_1$. The paper does not have an abstract;
EC$_2$. The paper is just published as an abstract;
EC$_3$. The paper is written in a language other than English;
EC$_4$. The paper is not a primary paper (e.g., keynotes);
EC$_5$. The paper is not accessible on the Web;

4) Data storage: The data extracted in the searching phase were stored into a spreadsheet that recorded all relevant details from selected papers. This spreadsheet supported classification and analysis procedures throughout this systematic mapping review.

5) Assessment: According to Kitchenham et al. [**?**], the consistency of the protocol utilized in a systematic mapping should be reviewed to confirm that: "the search strings, constructed interactively, derived from the research questions"; "the data to be extracted will properly address the research question(s)"; "the data analysis procedure is appropriate to answer the research questions". In this research, the first author is a Ph.D. candidate running the experiments. The second and third authors conducted the review process.

C. Data extraction and synthesis

Figure 3 illustrates the four-stage selection process carried out in the current mapping review. In each stage, the sample size was reduced based on the inclusion and exclusion criteria. In the first stage, relevant papers were retrieved by querying the databases with the search string presented in Table II. All database queries were issued in December 2017. This yielded a total of 54 initial papers: 13 from **IEEE Xplore**, 5 from **Science Direct**, 17 from **ACM Digital Library**, and 19 from **SpringerLink**. After removing four duplicates, we reduced (around 7%) the initial set of 50 papers. In the second stage, inclusion and exclusion criteria were applied over title, abstract, and keywords, reaching a set of 18 papers (reduction around 64%): 32 papers were rejected for not satisfying IC$_1$ (The paper discusses severity prediction of bug reports in FLOSS projects). In the third stage, exclusion criteria were applied considering the full text. However, no paper was rejected by taking into account these criteria.

In the fourth stage, the Snowballing (search for references) [**?**] activity was conducted, which resulted in 12 additional papers. After applying selection criteria over title, abstract, and keywords, 11 papers remained (reduction of 8.3% over the papers selected by snowballing). For these papers, selection criteria were applied considering the full text, and nine papers remained (reduction of approximately 18.2% over the 11 previously selected papers); EC$_3$ criterion eliminated one paper (the paper is written in a language other than English); EC$_6$ eliminated one more paper (the paper is not accessible on the Web), and another one was rejected for not satisfying IC$_1$. The selection phase ended up with 27 papers to be analyzed (18 from the sources plus nine from snowballing). Table IV shows the bibliographic reference of selected papers and an reference identifier (ID) for each paper. Throughout the remainder of this text, these identifiers will be used to refer to the corresponding paper.

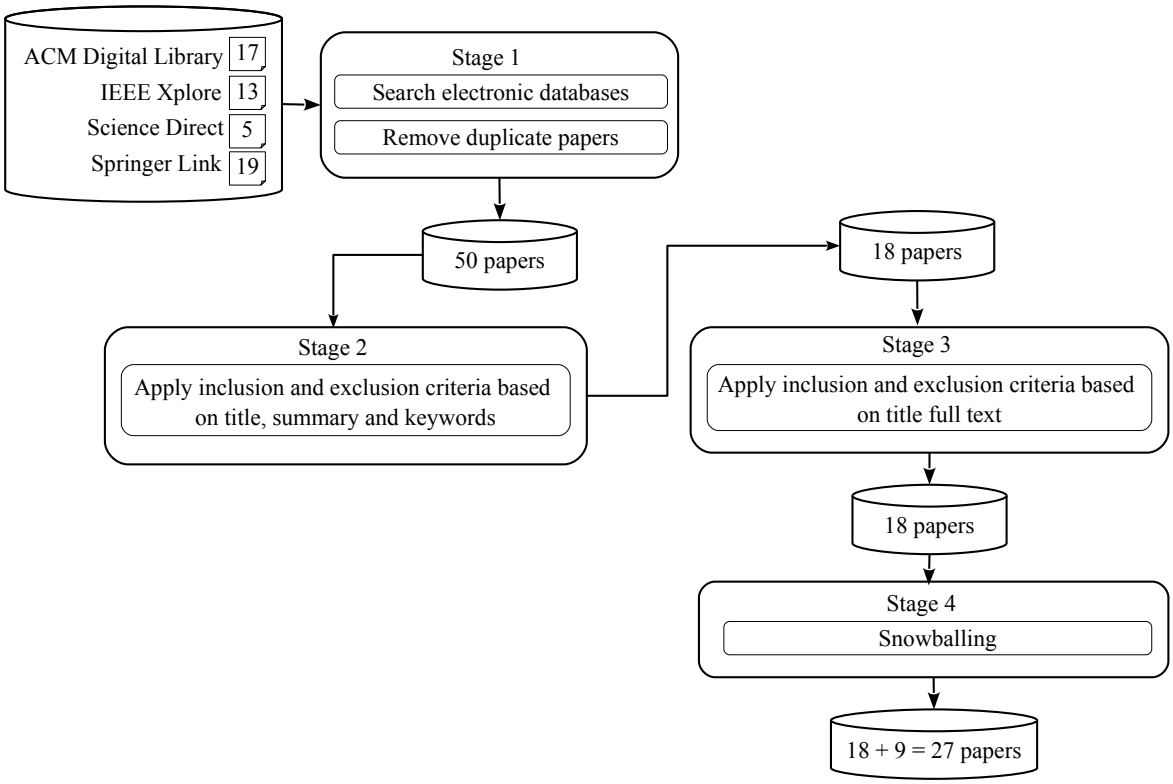

Fig. 3: Diagram of the four-stage paper selection process.

D. Limitations of this mapping

To interpret the implications of our results adequately, one needs to consider the following common limitations of all mapping reviews:

- **Mapping review completeness can never be guaranteed [?]**: Although this mapping review has observed a strict research protocol to ensure the relatively whole population of the relevant literature, some important papers might be missed.
- **Terminology problems in search string may lead to miss some primary studies [?]**: The terminology applied in the database query is normally accepted and used within the scientific community. Nevertheless, different terms may be used to retrieve the same relevant information.
- **Subjective evaluation of inclusion and exclusion criteria may cause misinterpretation [?]**. Explicit criteria have been defined for assessing the relevance of selected papers IV-B. However, the evaluation was based on the perception and experience of the authors. Different people may have other views regarding the relevance of these papers.

E. Categorization scheme

Petersen et al. [**?**] suggest the definition of a categorization scheme for conducting a systematic mapping analysis. The categories defined for this mapping review were based on: (i) existing labels commonly used in the literature and (ii) the content of selected papers. Next sections outline the categories considered in this mapping review.

1) FLOSS software type (RQ_2): This scheme organizes FLOSS projects by software type. Based on the taxonomy suggested by Pressman [**?**], FLOSS in studies belongs to three categories:

a. **Application software**: This category includes computer programs designed to solve a specific problem or business need for end users.
b. **System software**: This category includes collections of computer programs (operating systems and utilities) required to run and maintain a computer system.
c. **Programming tools**: This category includes computer programs that aid software developers in creating, debugging, maintain or perform any development-specific task.

TABLE IV: Selected papers

ID	Bibliographic reference
[**?**]	. doi:10.1109/MSR.2010.5463284.
[**?**]	. doi:10.1109/CSMR.2011.31.
[**?**]	. doi:10.1109/WCRE.2012.31.
[**?**]	. doi:10.1109/APSEC.2012.144.
[**?**]	. doi:10.4018/jossp.2012040103.
[**?**]	. doi:10.1109/ICSESS.2014.6933548.
[**?**]	. doi:10.1109/COMPSAC.2014.16.
[**?**]	. doi:10.1145/2597073.2597099
[**?**]	. doi:10.1007/978-3-319-09156-3_17
[**?**]	. doi:10.1109/SEAA.2014.51
[**?**]	. doi:10.1109/MSR.2015.31
[**?**]	. doi:10.1145/2695664.2695872
[**?**]	. doi:10.1016/j.procs.2015.10.059
[**?**]	. doi:10.1016/j.infsof.2014.12.006
[**?**]	. doi:10.1109/ABLAZE.2015.7154933
[**?**]	. doi:10.1109/CIMCA.2016.8053276
[**?**]	. doi:10.1109/IACS.2016.7476092
[**?**]	.
[**?**]	. doi:10.1007/s10664-015-9409-1
[**?**]	. doi:10.1016/j.jss.2016.02.034
[**?**]	. doi:10.14257/astl.2016.129.05
[**?**]	.
[**?**]	. doi:10.12988/ces.2016.6695
[**?**]	.
[**?**]	. doi:10.1145/3019612.3019788
[**?**]	. doi:10.1142/S0219649217500058
[**?**]	. doi:10.1109/SEAA.2017.71

2) Severity prediction problem (RQ_3): This scheme drills down severity prediction problems, based on studying of selected papers, in five categories:

a. **Severe or Non-Severe (SNS)**: This category comprises the prediction problems whose the predicted severity level might be severe or non-severe. In Bugzilla, for example, the severe class would include blocker, critical, and major; and the non-severe class would include minor and trivial levels. The predictors of this problem category do not take into account the default severity level.

b. **Severe or Non-Severe With Default Class (SNSWD)**: This category is quite similar to SNS. However, the predictors consider the default severity level as severe or non-severe, or yet as another class.

c. **Multiple Classes (MC)**: This category comprises prediction problems whose predictors treat each severity level as a different class. Likewise that in SNS, the predictors do not take into account the default severity level.

d. **Multiple Classes With Default Class (MCWD)**: This category is similar to MC. However, the predictors consider the default severity level as a regular class.

e. **Blocking or Non-Blocking (BNB)**: The prediction problem whose predictors classify a bug report severity level into blocking or non-blocking class. A blocking bug is a software defect that prevents other defects from being fixed [**?**].

3) Feature data types (RQ_4): This scheme groups the features used for bug reports severity prediction into five categories:

a. **Qualitative categorical**: This category encompasses features that contain an unordered list of values. For instance, bug report *platform* attribute (e.g., “Windows”, “Linux”, “Android”).

b. **Qualitative ordinal**: This category encompasses features that contain an ordered list of values. For instance, bug report *priority* attribute (e.g., P1, P2, P3).

c. **Quantitative discrete**: This category encompasses features that contain a finite or an infinite number of values that can be counted, such as the *bug fixed time* attribute.

d. **Quantitative continuous**: This category encompasses features that contain an infinite number of values that can be measured, such as the *summary weight* attribute.

e. **Unstructured text**: This category encompasses features that either do not have a pre-defined data model or are not organized in a pre-defined manner. For instance, the bug report *description* attribute usually includes free format texts.

4) Feature selection methods (RQ_5): This scheme organizes the feature selection methods used for bug reports severity prediction into three categories, according to the taxonomy described in Guyon and Elisseeff [**?**]:

a. **Filter**: This category comprises methods that assign a score to the features, and use this value as a selection criterion to identify the top-scoring ones.
b. **Wrapper**: This category comprises methods that prepare, evaluate and compare feature combinations, and select the best one. Methods within this category consider the feature selection as a search problem.
c. **Embedded**: This category comprises methods that select which features best contribute to the accuracy while the ML algorithm creates the predicting model.

5) Text mining feature representations (RQ_6): This scheme organizes the techniques for text mining feature representations used for bug report severity prediction in four categories. The following categories were employed in this systematic mapping review according to [**?**]:

a. **Character**: This category includes techniques that represent features as individual component-level letters, numerals, special characters, and spaces are the building blocks of higher-level semantic features, such as words, terms, and concepts.
b. **Word**: This category includes techniques that represent features as specific words selected directly from a "native" document.
c. **Term**: This category includes techniques that represent features as single words and multiword phrases selected directly from a corpus of a native document using a term-extraction method.
d. **Concept**: This category includes techniques that represent features generated for a document employing manual, statistical, rule-based, or hybrid categorization methodologies.

6) ML algorithms categories (RQ_7): This scheme groups ML algorithms used in bug report severity prediction in five categories. The following categories were based on taxonomy proposed by Facelli et al. [**?**]:

a. **Distance-based**: This category comprises algorithms that use the proximity between data to generate their predictions.
b. **Probabilistic-based**: This category comprises algorithms that make their predictions based on the Bayes's theorem.
c. **Searching-based**: This category comprises algorithms that rely on searching for a solution in space to generate their predictions.
d. **Optimization-based**: This category comprises algorithms whose predictions are based on an optimization function.
e. **Ensemble-method**: This category comprises algorithms that combine or aggregate results of different base classifiers to make a prediction.

7) Evaluation measures categories (RQ_8): This scheme organizes evaluation measures used in selected papers by categories. According to Japkowicz and Shah [**?**], there are six categories:

a. **Single class focus**: This category comprises measures based on confusion matrix information, whose focus is on a single class of interest.
b. **Multi-class focus**: This category comprises measures based on confusion matrix information, whose focus is on all the classes of the problem domain.
c. **Graphical measure**: This category comprises measures based on a confusion matrix in conjunction with extra information. They enable visualization of the classifier performance under different skew ratios and class priors distribution and are indicated for scoring classifiers.
d. **Summary statistics**: This category comprises measures based on a confusion matrix in conjunction with extra information. They enable to quantify the comparative analysis between classifiers and are indicated for scoring classifiers.
e. **Distance/error-measure**: This category comprises measures based on a confusion matrix in conjunction with extra information. They measure the distance of an instance's predicted class label to its actual label and are recommended for continuous and probabilistic classifiers.
f. **Information theoretic measures**: This category comprises measures based on a confusion matrix in conjunction with extra information. They reward a classifier upon correct classification relative to the (typically empirical) prior to the data and are indicated for continuous and probabilistic classifiers.

8) Sampling techniques (RQ_9): This scheme organizes the sampling techniques used for bug report severity prediction into three categories. Such categories according to Japkowicz and Shah [**?**] include:

a. **No re-sampling**: This category encompasses techniques that test the algorithm on a large set of unseen data.
b. **Simple re-sampling**: This category encompasses techniques that use each data point for testing only once.
c. **Multiple re-sampling**: This category encompasses techniques that use each data point for testing more than once.

9) Statistical Tests (RQ_{10}): This scheme groups statistical tests commonly used for bug report severity prediction in three categories. Such categories employed in this mapping review according to Japkowicz and Shah [**?**] include:

a. **Parametric**: This category comprises methods that make strong assumptions about the distribution of the population
b. **Non-parametric**: This category comprises methods that do not make strong assumptions about the distribution of the population

c. **Parametric and non-parametric**: This category comprises methods that are both parametric and non-parametric

10) Experiment software tools (RQ_{11}): This scheme groups the software tools used in experiments for bug reports severity prediction in two well-known and popular categories:

a. **Free/Libre Open Source Software (FLOSS)**: This category includes software that can be freely used, modified, and redistributed.
b. **Closed Source Software (CSS)**: This category includes software that is owned by an individual or a company whose source code is not shared with the public for anyone to look at or change.

V. Results

This section summarizes the results of this mapping review. The answer to each research question (from RQ1 to RQ12) is presented in tables and charts. The organization of extracted data follows the criteria defined in Section IV.

A. When and where have the studies been published? (RQ1)

Figure 4a shows that research on bug report severity prediction in FLOSS is recent and active with a vast number of papers (22 out of 27 $\approx$ 81%) published from 2014. Journals and conferences on information technology, mining software repositories, and software engineering seem to be more open to papers of bug report severity prediction (Table V).

TABLE V: Papers publication sources.

Paper source	Type	Reference
ACM Symposium on Applied Computing	Conference	[**?**]
Advanced Science and Technology Letters & Journal	Journal	[**?**]
Asia-Pacific Software Engineering Conference	Conference	[**?**]
Computational Science and Its Applications (ICCSA)	Conference	[**?**]
Computer Software and Applications Conference	Conference	[**?**]
Contemporary Engineering Sciences	Journal	[**?**]
Empirical Software Engineering	Journal	[**?**]
Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)	Conference	[**?**]
Information and Software Technology	Journal	[**?**]
International Conference on Circuits, Controls, Communications and Computing (I4C)	Conference	[**?**]
International Conference on Computer Science and Software Engineering	Conference	[**?**]
International Conference on Information and Communication Systems (ICICS)	Conference	[**?**]
International Conference on Software Engineering and Service Science	Conference	[**?**]
International Information Institute	Journal	[**?**]
International Journal of Open Source Software and Process(IJOSSP)	Journal	[**?**]
International MultiConference of Engineers and Computer Scientists	Conference	[**?**]
Journal of Information & Knowledge Management	Journal	[**?**]
Journal of Systems and Software	Journal	[**?**]
Procedia Computer Science	Journal	[**?**]
Software Engineering and Advanced Applications (SEAA)	Conference	[**?**], [**?**]
Symposium on Applied Computing (SAC)	Conference	[**?**]
Working Conference on Mining Software Repositories (MSR)	Conference	[**?**], [**?**], [**?**], [**?**]
Working Conference on Reverse Engineering	Conference	[**?**]

B. What FLOSS are the most used as experimental target for bug report severity prediction (RQ2)?

Table VI shows that most papers concentrated their focus on five FLOSS: Eclipse ($\approx$ 92%), Mozilla ($\approx$ 70%), Openoffice ($\approx$ 18%), Netbeans ($\approx$ 14%), and Gnome ($\approx$ 11%). The detailing of each FLOSS including description, category, BTS, and URL is presented in Table VII.

Figure 4b shows that most papers worked with FLOSS related to Programming Tool ($\approx$ 92%) and Application Software ($\approx$ 74%) categories. Moreover, it presents that 17 out of 27 ($\approx$ 62%) papers worked with both categories. Figure 4c highlights that all papers handled bug reports extracted from Bugzilla, and a just few papers from Google (3 out of 27 - $\approx$ 11%) and Jira (2 out of 27 $\approx$ 7%). Only one paper (Yang et al. [**?**]) investigated bug reports from three BTS.

TABLE VI: Paper distribution by FLOSS.

Project	References	Total
Eclipse	[?], [?]	25
Mozilla	[?], [?]	19
Openoffice	[?], [?], [?], [?], [?]	5
Netbeans	[?], [?], [?], [?]	4
Gnome	[?], [?], [?]	3
Chromium	[?], [?]	2
Freedesktop	[?], [?]	2
GCC	[?]	1
Hibernate	[?]	1
Spring	[?]	1
Android	[?]	1
JBoss	[?]	1
Mongo-db	[?]	1
WineHQ	[?]	1

TABLE VII: FLOSS investigated in papers.

Project	Description	Category	BTS	URL (As of September 10, 2018)
Android	Mobile operating system	System software	Google	https://issuetracker.google.com
Chromium	Web browser	Application software	Google	https://www.chromium.org/issue-tracking
Eclipse	Integrated Development Environment(IDE)	Programming tool	Bugzilla	https://bugs.eclipse.org/bugs/
FreeDesktop	Base platform for desktop software on Linux and UNIX	Programing Tool	Bugzilla	https://bugs.freedesktop.org/
GCC	C compiler	Programming tool	Bugzilla	https://gcc.gnu.org/bugzilla/
Gnome	Desktop environment based on X Windows System	Application Software	Bugzilla	https://gcc.gnu.org/bugzilla/
Hibernate	Object Relation Mapper(ORM) framework	Programming tool	Jira	https://hibernate.atlassian.net
Jboss	Application server	System software	Jira	https://issues.jboss.org/s
Mongo-db	No-sql database	System software	Jira	https://jira.mongodb.org/
Mozilla	Internet tools	Application software	Bugzilla	https://bugzilla.mozilla.org/
Netbeans	Integrated Development Environment(IDE)	Programming tool	Bugzilla	https://netbeans.org/bugzilla/
OpenOffice	Office suite	Application software	Bugzilla	https://bz.apache.org/ooo/
Spring	JEE Framework	Programming tool	Jira	https://jira.spring.io
WineHQ	Compatibility layer	System software	Bugzilla	https://bugs.winehq.org/

C. Was bug report severity prediction most addressed as either a fine-grained label or coarse-grained label prediction problem (RQ3)?

Table VIII shows that about the same fraction of papers addressed the bug report severity prediction as coarse-grained and fine-grained problem. Figure 5a shows that most papers (10 out of 27 $\approx$ 37%) addressed the severity prediction as a problem of SNS category. Moreover, 8 out of 27 ($\approx$ 20%) papers addressed it as a problem of the MC category and 5 out of 27 ($\approx$ 18.5%) addressed a problem of the MCWD category. A few papers addressed a BNB (2 out of 27 $\approx$ 7%) or an SNSWD (2 out of 27 $\approx$ 7%) problem. No paper addressed more than one problem category.

TABLE VIII: Paper distribution by prediction problem.

Problem	References	Total
Coarse-grained	[?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?]	14
Fine-grained	[?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?]	13

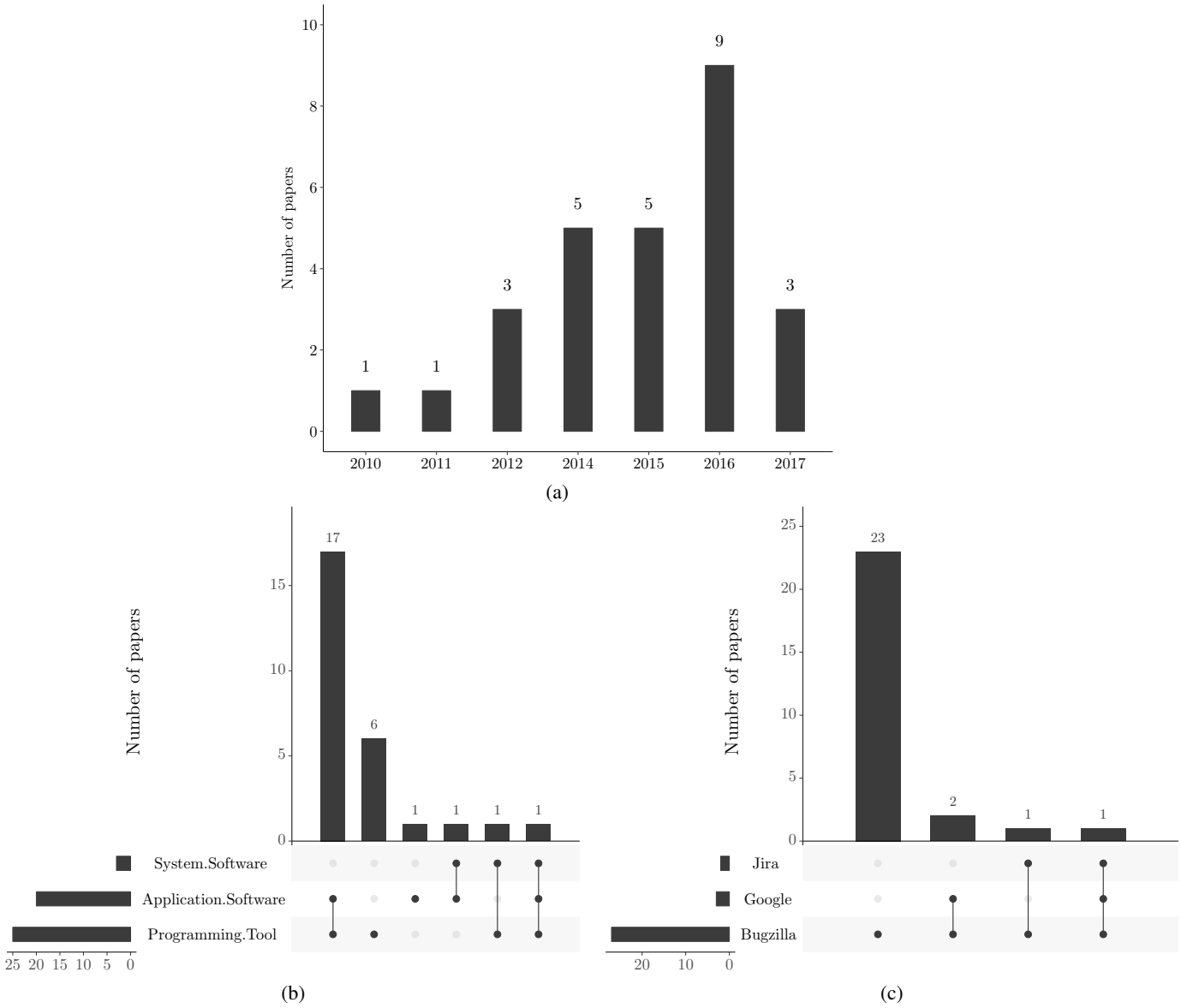

Fig. 4: (a) Paper distribution by year (67% published in conferences and 33% in journals), (b) paper distribution by FLOSS (c) paper distribution by BTS.

D. What are the most common features used for bug report severity prediction (RQ4)?

Table IX describes the features used for bug report severity prediction. Each feature in this table has a description, a category, an indicator of if the feature is computed from others, and an indicator of if the feature is available in Bugzilla, Jira, and Google.

As shown in Table X, most used features were *summary* ($\approx$ 74%) and *description* ($\approx$ 70.37%). Besides, a significant number of papers also reported using the features *product* ($\approx$ 37%) and *component* ($\approx$ 33%). Regarding data type of the features, unstructured text (26 out of 27 $\approx$ 96%) was most used data type in the papers (Figure 5b). Only one paper (Meera et al. [**?**]) reported the use of features belonging to each data type presented in the chart.

E. What are the most common features selection methods employed for bug report severity prediction (RQ5)?

Table XI shows that few papers ($\approx$ 40%) use of a feature selection method for bug report severity prediction and remarkably all papers only used filter category methods.

[1]In Jira, *product* corresponds to *project* field.

TABLE IX: Feature descriptions.

Feature	Description	Category	Calculated?	Bugzilla	Jira	Google
Attachments	Files (e.g., test cases or patches) attached to bugs.	Qualitative Categorical	No	Yes	No	Yes
Bug fix time	Time to fix a bug (Last Resolved Time - Opened Time).	Quantitative Continuous	Yes	Yes	Yes	Yes
Bug id	Bug report identifier.	Quantitative Discrete	No	Yes	Yes	Yes
CC list	A list of people who get mail when the bugs changes.	Qualitative Categorical	No	Yes	No	Yes
Comment	Textual content appearing in the comments of bug report.	Unstructured Text	Yes	Yes	Yes	Yes
Comment size	The number of word of all comments of a bug.	Quantitative Discrete	Yes	Yes	Yes	Yes
Complexity	Bug level of complexity based on bug fix time.	Qualitative Ordinal	Yes	Yes	Yes	Yes
Component	Each product is divided into different components (e.g., Core, Editor, and UI).	Qualitative Categorical	No	Yes	Yes	Yes
Description	Textual content appearing in the description field of the bug report.	Unstructured Text	No	Yes	Yes	Yes
Description size	The number of words in the description.	Quantitative Discrete	Yes	Yes	Yes	Yes
Importance	The importance of a bug is a combination of its priority and severity.	Qualitative Discrete	No	Yes	No	No
Number in CC list	The number of developers in the CC list of the bug.	Quantitative Discrete	Yes	Yes	No	Yes
Number of comments	Number of comments added to a bug by users.	Quantitative Discrete	Yes	Yes	Yes	Yes
Number of dependents	Number of dependents of a bug report.	Quantitative Discrete	Yes	Yes	Yes	Yes
Number of duplicates	Number of duplicates of a bug report.	Quantitative Discrete	Yes	No	No	Yes
Platform	Indicates the computing environment where the bug was found (e.g., Windows, GNU/Linux, and Android).	Qualitative Categorical	No	Yes	Yes	Yes
Priority	Priority should normally be set by the managers, maintainers or developers who plan to work, not by the one filling the bug or by outside observers.	Qualitative Ordinal	No	Yes	Yes	Yes
Product	What general "area" the bug belongs to (e.g., Firefox, Thunderbird, and Mailer).	Qualitative Categorical	No	Yes	Yes[1]	No
Report length	The content length of long description providing debugging information.	Quantitative Discrete	Yes	Yes	Yes	Yes
Reporter	The account of the user who created the bug report.	Qualitative Categorical	No	Yes	Yes	Yes
Reporter blocking experience	Counts the number of blocking bugs filed by reporter previous to this bug.	Quantitative Discrete	Yes	Yes	Yes	No
Reporter experience	Counts the number of previous bug reports filed by the reporter.	Quantitative Discrete	Yes	Yes	Yes	No
Reporter name	Name of the developer or user that files the bug	Qualitative Categorical	No	Yes	Yes	No
Severity	Indicates how severe the problem is – from blocker ("application unusable") to trivial ("minor cosmetic issue").	Qualitative Ordinal	No	Yes	Yes	No
Summary	A one-sentence summary of the problem.	Unstructured Text	No	Yes	Yes	Yes
Summary weight	Score calculated using the information gain criterion.	Quantitative Continuous	Yes	Yes	Yes	Yes

F. What are the most used text mining methods for bug report severity prediction (RQ6)?

As shown in Table XII, $\approx$ 74% out of the papers used unigrams for feature vector extraction and $\approx$ 33% out of the papers used TF-IDF for feature vector weighting. Regarding similarity/dissimilarity functions, $\approx$ 7% reported using some function: one paper used BM25, one paper BM25ext, and another KL divergence.

Figure 5c shows that most used text mining method (9 out of 15 $\approx$ 44%) belongs to term category. Besides, only one paper (Yang et al. [**?**]) reported the use of text mining belonging to each category presented in the chart.

G. What are the most used machine learning algorithms for bug report severity prediction (RQ7)?

Table XIII presents that most papers used k-NN and NB ($\approx$ 44%) and that NBM has also been used quite frequently ($\approx$ 40%). As shown in Figure 5d, the majority of the papers used a probabilistic-based algorithm (21 out of 27 $\approx$ 77%). The figure also shows that 15 out of 27 papers ($\approx$ 55%) used algorithms belonging to a ML algorithm single category.

[2]Topic Model [**?**] is a statistical model to identify "topics" from a collection of documents. Each topic includes the topic terms which appear in the documents and each document may belong to one ou more topics.

[3]BM25ext [**?**] is a extension of similarity function BM25. While BM25 computes the similarity of a short document with a long document, BM25ext computes the similarity between long documents.

[4]Kullback-Leibler(KL) [**?**] divergence measures the difference between two probability over the same variable.

TABLE X: Paper distribution by features.

Feature	References	Total
Summary	[?], [?]	20
Description	[?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?]	19
Product	[?], [?], [?], [?], [?], [?], [?], [?], [?], [?]	10
Component	[?], [?], [?], [?], [?], [?], [?], [?], [?]	9
Severity	[?], [?], [?], [?], [?]	5
Priority	[?], [?], [?], [?]	4
Reporter	[?], [?], [?]	3
Description size	[?], [?]	2
Number in CC list	[?], [?]	2
Reporter blocking experience	[?], [?]	2
Reporter experience	[?], [?]	2
Reporter name	[?], [?]	2
Attachments	[?]	1
Bug fix time	[?]	1
Bug Id	[?]	1
CC list	[?]	1
Comment	[?]	1
Comment size	[?]	1
Complexity	[?]	1
Importance	[?]	1
Number of comments	[?]	1
Number of dependents	[?]	1
Number of duplicates	[?]	1
Platform	[?]	1
Report length	[?]	1
Summary weight	[?]	1

TABLE XI: Paper distribution by feature selection methods.

Method	Category	References	Total
Information Gain	Filter	[?], [?], [?], [?], [?]	5
Chi-Square	Filter	[?], [?], [?]	3
Correlation Coefficient	Filter	[?]	1

[5]Näive Bayes Multinomial (NBM) [?] is similar to Näive Bayes. However, the output class is not only determined by presence or absence of term, but NBM also uses the number of occurrences of the terms to decide.

[6]C4.5 [?] is an algorithm used to generate a decision tree which follows a greedy divide and conquer strategy in training step.

[7]Bagging Ensemble [?] uses multiple learning algorithms to obtain better predictive performance that could be received from any of the constituent learning algorithms alone. This approach involves having each model in the ensemble vote with equal weight. To promote model variance, bagging trains each model in the ensemble using a randomly drawn subset of the training set.

[8]AdaBoost [?] is an ensemble algorithm that attempts to produce a very accurate classification rule by combining moderately inaccurate weak classifiers.

[9]Functional Tree [?] is a classification tree that could have logistic regression function at the inner nodes and or leaves.

[10]Random Tree [?] is an ensemble classifier that consists of directed graphs build with a random process.

[11]Radial Basis Function (RBF) Neural Network [?] is a neural network that consists of input nodes connected by weights to a set of RBF neuron, which fire proportionally to the distance between the input and the neuron in weight space. RBF is a real-valued function whose value depends only on the distance from the origin.

TABLE XII: Paper distribution by text mining models.

Technique	Category	References	Total
		Vector Extraction Models	
Unigrams	Character	[?], [?]	20
Bigrams	Character	[?], [?], [?], [?]	4
Topic model[2]	Concept	[?], [?], [?]	3
		Vector Weighting Models	
TF-IDF	Term	[?], [?], [?], [?], [?], [?], [?], [?], [?]	9
TF	Term	[?], [?], [?]	3
Binary	Term	[?]	1
		Vector Similarity/Dissimilarity Functions	
BM25ext[3]	Term	[?]	1
BM25	Term	[?]	1
KL divergence[4]	Term	[?]	1

TABLE XIII: Papers distribution by ML algorithms

Algorithm	Category	References	Total
KNN	Distance	[?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?]	12
NB	Probabilistic	[?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?]	12
NBM[5]	Probabilistic	[?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?]	11
SVM	Otimization	[?], [?], [?], [?], [?], [?], [?]	7
Random Forest	Ensemble	[?], [?], [?], [?], [?]	5
C4.5[6]	Searching	[?], [?], [?]	3
Bagging	Ensemble	[?], [?]	2
Decision Tree	Searching	[?], [?]	2
AdaBoost	Ensemble	[?]	1
Functional Tree	Searching	[?]	1
Random Tree	Ensemble	[?]	1
RBF Networks	Optimization	[?]	1
RIPPER[7]	Other	[?]	1
Zero-R[8]	Other	[?]	1

H. What are the measures used to evaluate ML algorithms performance for bug report severity prediction (RQ8)?

Most papers evaluated the ML model accuracy in bug report severity prediction using three measures based on the confusion matrix (Table XIV): precision ($\approx$ 77%), recall ($\approx$ 70%), and f-measure ($\approx$ 66%). The majority of the used measures are of the single class category (21 out of 27 $\approx$ 70%), as shown in Figure 6a. Only one paper (Roy et al. [**?**]) investigated all four categories of measures presented in the chart.

I. Which sampling techniques are applied most frequently to generate more reliable predictive performance estimates in severity prediction of a bug report (RQ9)?

Most papers shown in Table XV applied 10-Fold CV ($\approx$ 66%) to generate more reliable effectiveness estimates. Figure 6b displays that the majority of those papers (14 out of 15 $\approx$ 93%) employed a simple resampling method. Also, only one paper

[12]RIPPER [**?**] is a rule-based classifier that builds a set of rules that identify the classes while minimizing the amount of error. The error is defined by the number of training examples misclassified by the regulations.

[13]Zero-R [**?**] is the simplest classifier which always predicts the majority class in training set. It is commonly employed for determining a baseline performance as a benchmark for other methods.

[14]Mean Reciprocal Rank (MRR) [**?**] is the average of reciprocal ranks of results of a set of questions. A reciprocal rank of a question is the multiplicative inverse of the rank of the first correct answer.

[15]Effectiveness Ratio [**?**] is a cost-effectiveness measure, which evaluates prediction performance given a cost limit.

[16]Krippendorff's Alpha Coefficient [**?**] is a statistical measure of the agreement achieved when coding a set o units analysis of values of a variable.

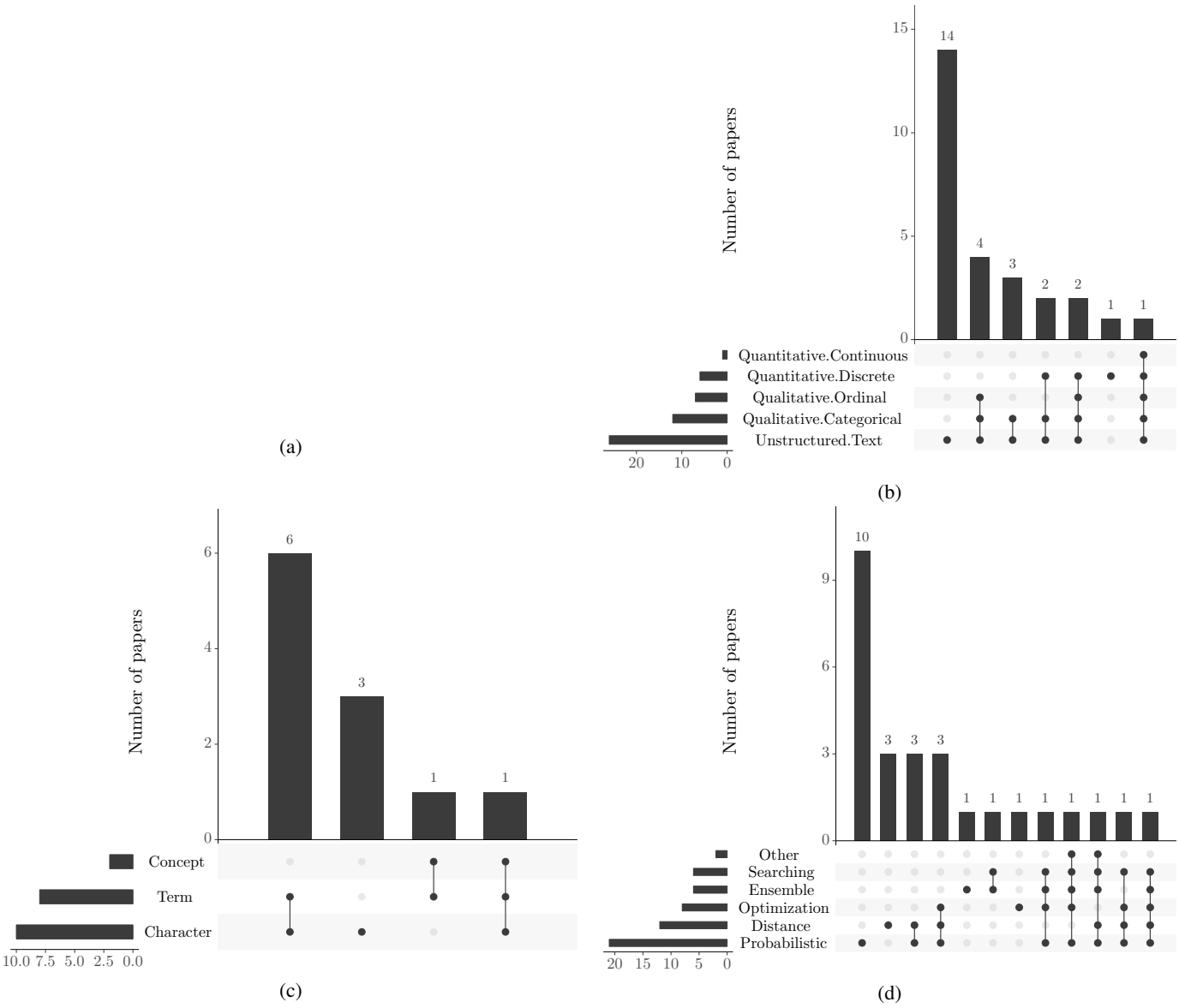

Fig. 5: (a) Paper distribution by prediction problem category, (b) paper distribution by feature data category, (c) paper distribution by text mining method category (d) paper distribution by ML algorithm categories.

TABLE XIV: Papers distribution by evaluation measures.

Measure	**Category**	**References**	**Total**
Precision	Single Class	[?], [?]	21
Recall	Single Class	[?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?]	19
F-measure	Single Class	[?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?]	18
Accuracy	Multi-Class	[?], [?], [?], [?], [?], [?], [?], [?], [?], [?], [?]	11
AUC	Summary	[?], [?], [?], [?], [?]	5
ROC	Graphical	[?], [?]	2
MRR[9]	Single Class	[?], [?]	2
Effectiveness Ratio[10]	Single Class	[?]	1
Krippendorff's Alpha Reliability[11]	Single Class	[?]	1

(Otoom et al. [**?**]) investigated both non-resampling and simple resampling methods for bug report severity prediction.

TABLE XV: Paper distribution by sampling methods.

Method	Category	References	Total
10-Fold CV	Simple Re-sampling	[**?**], [**?**], [**?**], [**?**], [**?**], [**?**], [**?**], [**?**], [**?**], [**?**]	10
Stratified 10-Fold CV[12]	Simple Re-sampling	[**?**], [**?**], [**?**]	3
SMOTE[13]	Simple Re-sampling	[**?**], [**?**]	2
Hold-out	No Re-sampling	[**?**]	1
03-Fold CV	Simple Re-sampling	[**?**]	1
05-Fold CV	Simple Re-sampling	[**?**]	1

J. Which statistical tests were used to compare the performance between two or more ML algorithms for bug report severity prediction (RQ10)?

As shown in Table XVI, most of the papers used Wilcoxon Signed Rank Test (8 out of 27 $\approx$ 29%) and t-test (7 out of 27 - $\approx$ 26%) to compare the performance of ML algorithms. Figure 6c calls attention that only ten papers ($\approx$ 37%) reported using any statistical test in their experiments to bug report severity prediction. The figure also shows that most of these papers (8 out of 10 $\approx$ 80%) applied a non-parametric test, and 5 out of 10 (50%) papers employed both parametric and non-parametric tests.

TABLE XVI: Paper distribution by statistical tests.

Test	Category	References	Total
Wilcoxon signed rank test	Non-parametric	[**?**], [**?**], [**?**], [**?**], [**?**], [**?**], [**?**], [**?**]	8
T-test	Parametric	[**?**], [**?**], [**?**], [**?**], [**?**], [**?**], [**?**]	7
Proportion test	Parametric	[**?**]	1
Shapiro-Wilk test	Parametric	[**?**]	1

K. Which software tools were used to run bug report severity prediction experiments (RQ11)?

Most of the papers listed in Table XVII used RapidMiner (5 out of 18 $\approx$ 27%) to run their experiments. However, Figure 6d presents a strong predominance of FLOSS over CSS, 13 out of 18 papers ($\approx$ 61%) executed FLOSS to perform their experiments. The figure also shows that no paper used both CSS and FLOSS.

TABLE XVII: Paper distribution by software tools.

Tool	Category	References	Total
RapidMiner	CSS	[**?**], [**?**], [**?**], [**?**], [**?**]	5
WEKA	FLOSS	[**?**], [**?**], [**?**], [**?**]	4
R	FLOSS	[**?**], [**?**], [**?**]	3
NLTK	FLOSS	[**?**], [**?**], [**?**]	3
Statisca	CSS	[**?**], [**?**]	2
TMT	FLOSS	[**?**], [**?**]	2
OpenNLP	FLOSS	[**?**]	1
WordNet	FLOSS	[**?**]	1
Ruby	FLOSS	[**?**]	1
WVTool	FLOSS	[**?**]	1
CoreNLP	FLOSS	[**?**]	1

[17]Stratified k-fold CV [**?**] is a k-fold CV variation, which ensures that class distribution is respected in the training and testing sets created at every fold.

[18]Synthetic Minority Oversampling TEchnique (SMOTE) [**?**] is an sampling approach in which the minority class is over-sampled by creating "synthetic" examples rather than by over-sampling with replacement.

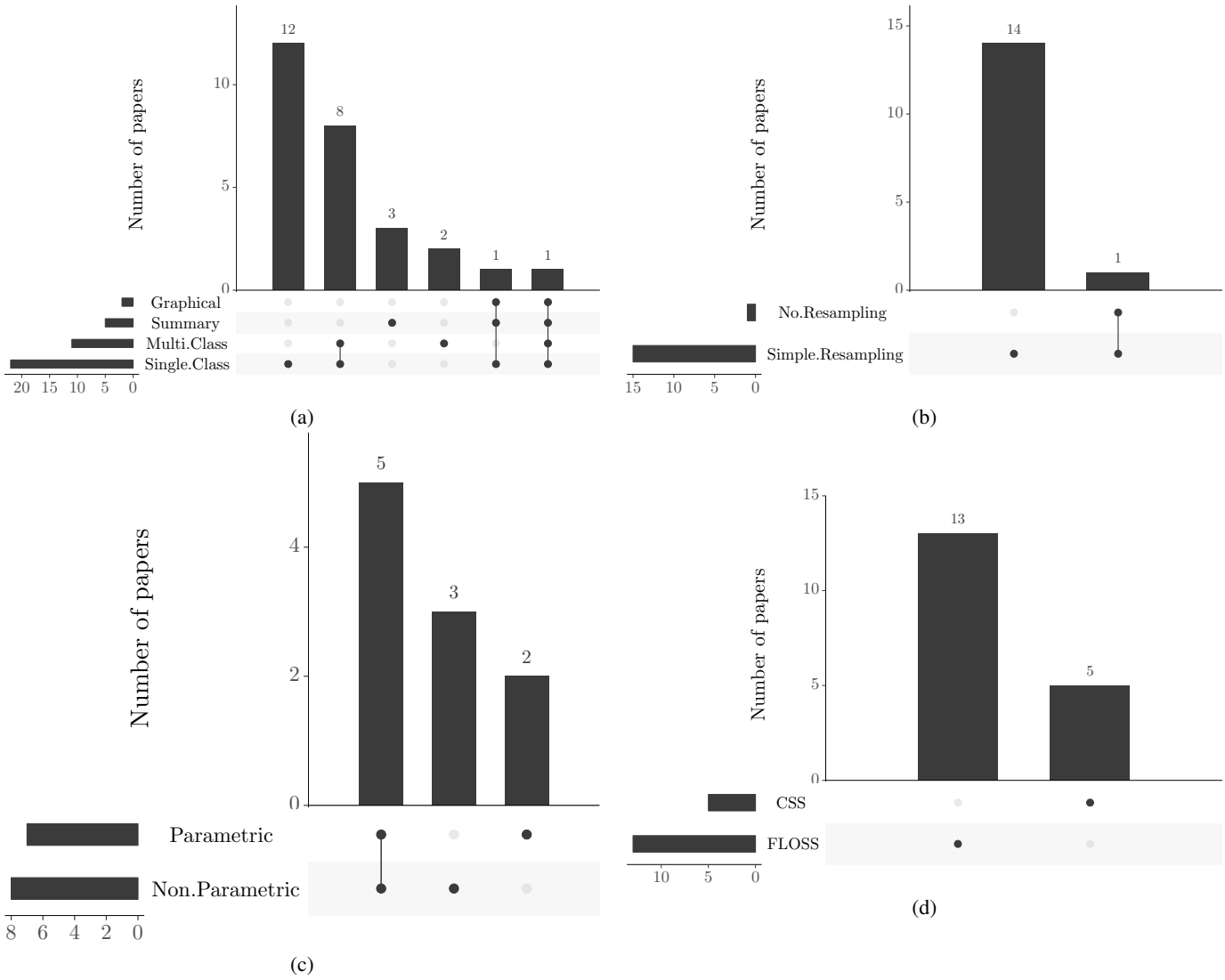

Fig. 6: (a) Paper distribution by evaluate measure categories, (b) paper distribution by sampling method categories, (c) paper distribution by statistical test categories, and (d) paper distribution by software tool categories.

L. Which solution types were proposed for bug report severity prediction problem (RQ12)?

Most of the papers (26 out of 27 ≈ 93%) provided offline solutions (Table XVIII). Just one paper (Tian et al. [**?**]) presented an online solution.

TABLE XVIII: Solution types proposed in selected papers.

Tool	References	Total
offline	[**?**], [**?**]	26
online	[**?**]	1

VI. Discussion

This section presents lessons learned from this systematic mapping study and recommends possible research directions for the bug severity prediction problem.

Researchers have concentrated their focus on few FLOSS, namely Eclipse and Mozilla, both hosted by Bugzilla. The main reason given by most researchers is benchmarking their results with Lamkanfi et al. [**?**], [**?**], which are often considered the state-of-the-art in this research area. Notably, they have not considered relevant FLOSS projects like as Linux Kernel, Spark,

MySQL, among others. Moreover, published research only marginally investigated two other bug trackers mentioned in reviewed papers: Jira and Google Issue Tracker. The first one tracks bugs for more than 80% Apache Foundation's projects[18], and the last one tracks bugs for many internal and external Google Inc. projects[19]. Surprisingly, for example, no paper investigated Github, another outstanding collaborative website which hosted more than 67 million projects (including many FLOSS)[20].

From 14 FLOSS identified in this review, six are programming tools, four are system software, and four are application software. For former two, people who report bugs are typically technical users, and, for the latter, people are end-users. A bug report can be more or less detailed depending on who reports it [**?**] [**?**]. Therefore, the probability of writing more accurate bug reports is larger for technical users than end-users.

Researchers considered the bug report severity prediction using five different approaches: three of them (SNS, SNSWB, and BNB) predict the severity level from two or three class (coarse-grained). Two other approaches (MC and MCWD) seem to be more complex than problems of the previous ones because the severity level may be predicted from five or six classes (fine-grained).

From another aspect, two categories (SNSWD and MCWD) did not consider the default severity level (in Bugzilla, often "normal"). By doing so, they have argued that they would avoid noisy and unreliable data. Saha et al [**?**] confirmed that many bug reports in practice are not "normal" and this misclassification can really affect the accuracy of ML algorithms.

Most of the approaches proposed by researchers to predict bug report severity relied on unstructured text features (*summary* and *description*). Because of this, text mining techniques played, side by side with machine learning methods, an essential role in delivering appropriated solutions. Two other features (*product* and *component*) was also quite used in these approaches. That may indicate that bug reports collected from separate parts of a single FLOSS yield different ML algorithms outcomes.

Regarding available feature data types in bug reports, more than half of them are qualitative. SVM and Neural Networks, two popular and essential ML algorithms in modern machine learning [**?**], do not work with qualitative data [**?**]. These two facts bring an extra challenge to use qualitative features for bug report severity prediction. When the input data set has qualitative data, categorical, and ordinal values must be converted into numeric values before applying ML algorithms.

Despite a large number of features raised by the application of text mining activities (e.g., tokenizing, stop word removal and stemming) on *summary* and *description* and of its effectiveness in SNS and SNSWD problems [**?**], few papers employed feature selection methods for bug severity prediction. Furthermore, these papers notably only investigated filter feature selection methods. Filter-based approaches have two drawbacks [**?**]: (i) they do not take into account redundancy between features, and (ii) they do not detect dependencies between them.

Most papers reported the use of unigram for feature extraction. However, few papers explicitly report the use of another text ming methods. Among these papers, the most used method feature vector weighting was TF-IDF and the most used method for verifying text similarity was BM25.

It seems that researchers have adopted a conservative approach regarding the use of ML algorithms. Most papers applied at least one of the following well-known and traditional supervised algorithms: k-NN, Näive Bayes, and its extension Näive Bayes Multinomial.

Researchers evaluated the performance of ML algorithms in their proposed approaches using precision, recall, and f-Measure. Three common measures employed in ML arena, but which may strongly skew in the imbalanced scenario [**?**]. This characteristic is particularly critical in bug report repositories which are intrinsically imbalanced in practice [**?**]. To minimize such distortion, Tian et al. [**?**], instead, recommend measuring performance using inter-rater agreement based metrics, such as Cohen's Kappa [**?**] or Krippendorff's Alpha [**?**]. Despite its well-known issues in an imbalanced dataset [**?**], quite few papers still used accuracy to measure ML performance. AUC, an alternative used by few papers, seems more robust than accuracy in imbalanced data conditions. Like Kappa and Krippendorff's Alpha, AUC considers class distribution on performance evaluation of ML algorithms for bug report severity prediction.

Few papers reported the use of a resampling approach to improving the accuracy of ML algorithms. The papers only reported the use of simple resampling strategies, including, the most used, k-fold cross-validation. Only one paper used SMOTE, considered a "de facto" resampling method in imbalanced data scenario [**?**]. Japkowicz et el. [**?**] also suggest that experiments may achieve better results using multiple resampling methods, for example, repeated k-fold cross-validation.

It seems that most researchers compared the results yielded in their experiments with others observing only the means of measures applied. Few of them reported utilizing statistical tests to do that, which is a recommended practice in empirical software engineering [**?**]. Most used statistical test was parametric. This kind of test requires that samples follow a statistical distribution (e.g., normal), which is not guaranteed to occur in practice when comparing ML models [**?**]. Besides, no paper investigated Analysis of Variance (ANOVA), a robust ranked-based test recommended by Kitchenham et al. [**?**] for Empirical Sofware Engineering.

Researchers preferred to use FLOSS tools to perform their analyses. Although, many papers have used a proprietary software named RapidMiner. Interestingly, researchers demonstrated a low interest in top-ranked ML tools based on R and Python programming languages.

[18]http://www.apache.org/index.html#projects-list (As of September 2018).
[19]https://developers.google.com/issue-tracker/ (As of September 2018).
[20]https://octoverse.github.com/ (As of September 2018)

It seems that most proposed solutions for bug report severity prediction were conceived to run in offline mode, just one paper claimed which the published solution is mature and fast enough to be embedded, for example, in a web browser. The need for a high number of labeled bug reports during the training phase [**?**] is a problem intrinsic to supervised ML algorithms, which may make it difficult to turn these offline solutions to online.

VII. Conclusions and Research Directions

A systematic mapping review provides a structure for a research report type, which enables categorizing and giving a visual summary of results that have been published in papers of a research area [**?**]. This map aids to identify gaps in a research area, becoming a basis to guide new research activities [**?**]. The current mapping review captured the current state of research on bug report severity prediction, characterized related problems and identified the main approaches employed to solve them. These objectives were reached by conducting a mapping of existing literature. In total, the review identified 27 relevant papers and analyzed them along 12 dimensions. Although these papers have made valuable contributions in bug report severity prediction, the panorama presented in this mapping review suggests that there are potential research opportunities for further improvements in this topic. Among them, the following research directions appear to be more promising:

- There is an apparent lack of investigation on bug report severity prediction in other relevant FLOSS such as, for example, Linux Kernel, Ubuntu Linux, and MySQL, and in others BTS, for example, Github.
- Often, technical users report most bugs. Thus, the influence of user experience in predicting outcomes is still overlooked.
- Bug reports labeled with default severity level (often "normal") were prevalent in the most datasets used in reviewed papers. However, they are considered unreliable [**?**], and just discarding them also does not seem appropriate. Then, efforts in researching on novel approaches to handle this type of report should be considered to improve the state-of-the-art of severity prediction algorithms.
- Most approaches were based on unstructured text features (*summary* and *description*). To handle them, researchers chose to use the traditional bag-of-words approach instead of more recent text mining methods (e.g., word-embedding [**?**]) or data-driven feature engineering methods which may likely improve outcomes yielded so far.
- There is a clear research opportunity to investigate whether state-of-the-art ML algorithms might outperform the traditional algorithms used in all reviewed papers for bug report severity prediction. The investigation of the use of Deep learning algorithms which perform very well when classifying audio, text, and image data [**?**] seems to be a promising research direction.
- Researchers should investigate more recent techniques (e.g., continuous learning [**?**]) to provide an approach for bug report prediction which could be employed in real-world scenarios.
- Many bug reports are resolved in a few days (or in a few hours) [**?**]. Efforts to predict severity level for these group of bug reports do not seem very useful. Thus, an investigation to confirm this hypothesis and to determine when the severity prediction is more appropriate in bug report lifecycle is of critical importance.

Acknowledgment

Authors are grateful to CAPES (grant #88881.145912/2017-01), CNPq (grant #307560/2016-3), FAPESP (grants #2014/12236-1, #2015/24494-8, #2016/50250-1, and #2017/20945-0) and the FAPESP-Microsoft Virtual Institute (grants #2013/50155-0, #2013/50169-1, and #2014/50715-9).