IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol. 9, Issue.2, February 2021 E-ISSN: 2347-2693

An Overview of the State of Machine Learning in Bug Report Summarization

Som Gupta[1*], S.K. Gupta[2]

[1]Research Scholar AKTU Lucknow, UP, India
[2]Associate Professor, Computer Science Department BIET Jhansi, India

[*]*Corresponding Author: somi.11ce@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v9i2.5356 | Available online at: www.ijcseonline.org

Received: 10/Feb/2021, Accepted: 18/Feb/2021, Published: 28/Feb/2021

***Abstract*—** Bug Report is one of the most consulted software artifacts during the software evolution and maintenance process. Summarization is one of the approaches which is generally performed over them to perform Bug Report Analysis tasks like Duplicate Bug Report Analysis for Bug Triagers, Quick understanding of Bug Reports, Classification of Bug Reports into priorities, etc.
Information Retrieval Techniques, Natural Language Processing Techniques, Machine Learning Techniques and Deep Learning Based Techniques have been successfully implemented for doing the task. Machine Learning is one of the very popular techniques which has been used by almost 70 percent of the researchers for performing the Bug Report Summarization task.
Machine Learning is a very common technique which is used in context of Bug Report Summarization due to the fact that the Bug Reports are very domain-specific in nature .In this paper we have systematically analyzed the Machine Learning works used for Bug Report Summarization. We have chosen all the popular papers available through Springer, IEEE, ACM, ACL Anthology and Google Scholar.

***Keywords*—**Bug Report, Machine Learning, Supervised Learning, Unsupervised Learning, Classifiers

I. INTRODUCTION

With the advent of the internet, most of the organizations use the internet as a medium to store their artifacts. Be it the Mobile Companies, Online E-Commerce or any organization, they take the feedback and have many facilities for users for the product support. As with internet globalization, the number of users are usually very large. It is not possible to analyze all the information posted by users. So text summarization is used in these contexts.

Similarly for the Software projects, which are so many because everything is happening through the internet nowadays because of the capability of wider reach; text summarization is one very popular activity done for them.
In Software project development, a team goes through many phases before rolling out to the customers. These phases in broader terms include Requirement Engineering, Designing,Implementation and Testing. In all the phases because of the fact that the team is usually geographically dispersed, everything is documented and properly archived.

During all the phases, errors and defects are common. These are written in the form of Bug Reports.

A Bug Report is a software artifact generated during the whole software process. It is very important for an organization due to the fact that every organization wants their software to be competitive and bug free for easy access to users.

Bug Reports are stored in Software Bug Repositories. These Repositories also give the facility to any user over the world to report the bug if any is observed. Because of many users, there is a high probability that the same Bug is reported many times. Also because for every software, usually there will be a lot of roll outs, Regression testing is one which has to be performed very frequently. Thus for resolving an issue generated because of the implemented facility, the knowledge of previous similar Bug Report is very beneficial [1].

Bug Report Summarization is one of the very popular research topics while analyzing the Bug Reports. Even though Natural Language Processing, Machine Learning, Information Retrieval and Deep Learning techniques [2] have been applied to Bug Report summarization; Machine Learning is very popular . The reason is the less complex nature and ability to distinguish whether the sentence is important or not. Be it the extractive summarization or abstractive summarization on Bug Reports, first the important sentences are to be found [3]. Because of the domain-specific nature of these artifacts, feature analysis is performed in almost all the research.

The paper is organized as follows. Section 2 describes the various techniques available for Bug Report

Summarization. Section 3 briefly discusses the works in the field of Bug Report Summarization where the Machine Learning Techniques have been used. Section 4 briefly mentions the strengths and limitations of Machine Learning Techniques and finally the conclusion.

II. TECHNIQUES USED

- **CENTROID:** It is one of the popular unsupervised approaches to extract the labels where the centrality is the main focus. In order to further study about the centroid sentences, TF IDF is used [4].

- **GRASSHOPPER:** It is a graph based unsupervised approach. It is a variant of PageRank where the sentences of the text become the nodes of the graph and the relation between the sentences become the edges of the graph. The relation is represented by some similarity measure. In the Grasshopper based approach the absorbing random walks concept is used along with the Markov Chains. It is a diversity based approach [4].

- **MMR:**It is Maximum Marginal Relevance approach. It is one of the very famous approaches used for the multi-document summarization. It is a greedy approach and query specific approach. The aim is to find the sentences which have lesser similarity with the previously selected sentences and the maximum relevance to the query .

- **DECISION TREE:** It creates the decision tree with the data. There are two types of decision trees. One is Regression based and the other one is Classification based. Regression based trees are used when the continuous value has to be obtained. Whereas classification based trees are used when the categorical or discrete values like True or False, Important or Non Important needs to be obtained. For the Bug Report Summarization, as in most of the cases where the classifiers are used, the main idea is to find out whether the sentence is important for consideration to the summary generation.

- **NAIVE BAYES:** It is a classification algorithm which is a probabilistic classifier. It is based upon the assumption that all the variables in the feature vector are independent and equal. It finds the probability of happening of one event given the probability of happening of another event.

- **K-NEAREST NEIGHBOURS:** It uses the tree data structure. Its efficiency depends upon the number of neighbors requested and the number of features present for the classification. For choosing the neighbors either the Brute force approach or K- D Tree is used.

- **NEURAL NETWORKS:** Neural Networks also involve learning. Here the input layer, hidden layers and output layer are involved in it. This branch of machine learning which involves the neural networks with multiple hidden layers is also known as deep learning. And it is gaining a lot of popularity because of the fact that it captures the non-linear relations very nicely. Many variations like Convolutional Neural Networks, Recurrent Neural Networks, Gated Recurrent Neural Networks along with the attention mechanisms are used to improve the efficiency of neural networks.

- **SVM:** It is one of the supervised algorithms which is used for both the classification and regression based approach. The aim is to find out the line of separation or decision boundary which classify the n-dimensional data. They work on the fact that they try to find out the extreme points which are known as support vectors. For separating the linearly separable data, Linear SVM are used and for separating the non linearly separable data, Non-Linear SVM are used.

III. RELATED WORK

Machine Learning Approaches include supervised and unsupervised approaches. Supervised Machine Learning techniques involve the learning from the training data with the labels in it. While the unsupervised approaches do not require the labelled training data. Both the approaches have been used by various researchers for various purposes.
Machine Learning approaches have been successfully implemented for the task of Text Summarization for almost all the popular areas like Movies data, Scientific Research Journals data, Discussion sites, Books data, etc. They have also been used for the conversational informal data like email threads. In this paper, we focus on the summarization of Bug reports.

As the Bug Report is an informal and unstructured document involving multi-threaded information and a lot of technical information from various domains, the summarization is a big challenge for the research community. Even though for general text summarization, researches have been done for all the variations of text summarization, like extractive, abstractive, multi-document, etc. But because of the extreme domain-specificness and language handling issues with the Bug Reports, only extractive summarization is popular due to the fact that it is costly to create abstractive summaries. For summarization of Bug Reports, sentence-level summarization is the most popular among all. Not many researches have been conducted in this field. Out of all the research which has been conducted in the field of Bug Report Summarization, we have mentioned only research where the Machine Learning algorithms have been used.

Mani et al. [4] in their AUSUM approach do not use the Supervised Learning Approach to create the Bug Report Summaries. They identified the four Unsupervised Approaches namely MMR, PageRank, DivRank and GrassHopper and applied to the Bug Report Summarization to observe how they work when applied in context of Bug Reports.

Lotufo et al. [5] in their hurried way of approach used the concept of variation of PageRank to identify the patterns the users take while finding out the important sentences for consideration to the summary when they are in a hurried manner.

Rastkar et al. [6] [7] in the BRC based approach used the 24 features namely classified into Structural, Participant, Length and Lexical. They used the classifiers of Supervised Machine Learning Techniques to find out if the sentence is important or not.

Yang et al. [8] also used the approach similar to Rastkar et al. but extended the approach by using a noise filter to filter out the sentences which they thought are not important for the summary generation.

For the categorization of Bug Reports, clustering has been used by [9]. With this approach the bug reports were grouped together according to the text similarity. This technique can be very useful for multi-document summarization in Bug Reports. Huai in [10] used the intentions by using the supervised training approach to improve the Bug Report Summaries. They used seven intention categories and found the connection to the Bug Reports. The seven features are Bug Description, Fix Solution, Opinion Expressed, Information Seeking, Information Giving, Meta/Code, and Emotion Expressed. They used the same BRC dataset but they annotated it with the intentions. Along with the four sets of features used by Rastkar i.e structural features, participant features, length features, lexical features; they added one more set of features known as intention features.

IV. METHODOLOGY USED

In the paper, we have used the keyword based approach. We have included the papers from all the major libraries available for this field of research. Our work involves the use of ACL Anthology, Elsevier, Springer and ACM.

First we collected all the papers, involving the research work in the field of Bug Report Summarization. Then we read the abstract of all the papers and filtered the papers which involved the machine learning approach for their work.

We found that not many papers are available for the Bug Report Summarization and when we filtered we found very few papers were available. After filtering those few papers, we classified them according to the technique they used.

V. STRENGTHS AND LIMITATIONS

1. If we are using the supervised machine learning approach for learning, it requires a huge training set for good results. But for Bug Report Summarization, there are very few datasets which are available and their size is also not huge. This limits the testing of supervised learning approaches for Bug Report Summarization. Also it requires the labelled data and thus if we want to learn something which we do not know, it limits the usage of these approaches whereas in case of unsupervised learning approaches, they do not require the labelled data and thus helps in identifying the hidden information from the text.
2. Even though there are so many classifiers and approaches which have been developed and used successfully for a lot of applications, there is no algorithm which works the same for all the applications. For different applications, according to size and nature of the dataset, the efficiency of the algorithm changes.
3. It is easy to use supervised learning approaches whereas as the Unsupervised learning approaches do not require the training data, it is very difficult to understand what is happening inside these approaches.
4. As the data in the Bug Reports is unstructured, includes lot of noise and diversity, Machine Learning approaches alone cannot prove to be very beneficial as the noise and inclusion of domain-specific features, limit the efficiency of Machine Learning Approaches
5. Machine Learning approaches can easily interpret the data with multi-dimensional features. Also the information retrieval and natural language processing can be integrated into it to get the better results.
6. Unsupervised algorithms also suffer from underfitting and overfitting problems. Thus understanding of the dataset is very important before we apply any machine learning approach.

VI. CONCLUSION AND FUTURE SCOPE

Bug Report Summarization helps users find the relevant Bug Report quickly. Bug Report Summarization is a very challenging task due to the fact that the Bug Report is a very unstructured document which involves the multi-threaded communication among the user community. Bug Reports can involve discussion related to any type of software domain. Thus constructing a generic technique which considers the domain knowledge to find out the essence of the Bug Report is very difficult.

Machine Learning Approaches are one of the popular approaches which have been used in combination with the other approaches for Bug Report Summarization for the reason that the features help capture the domain knowledge. With the advancement of technology and research in the field of Machine Learning, there are many classifiers which help capture the non-linear type of data also nicely. Unsupervised , Supervised and Neural Networks all have been used by different researchers for their different approaches.

Even though there are many classifiers and many complex network based unsupervised approaches which are available for many tasks but very few classifiers have been tried for Bug Report Summarization. Testing the summarization by including various classifiers which have been used for generic text summarization will help

understand how the Machine Learning Approaches will work with the Domain-Specific artifact. Also as the Bug Report resembles the Email Threads, understanding the working in context of Bug Reports will supplement the information for Emails also.

REFERENCES

[1] Barzilay, R., & McKeown, K. R. "Sentence fusion for multidocument news summarization." Computational Linguistics, **vol 31, pp. 297–327, 2005.**

[2] Gupta, S., & S.K, G. "Deep learning in automatic text summarization." International Journal of Computer Science and Information Security (IJCSIS), **vol. 16, pp. 150–155, 2018.**

[3] Gupta, S., & Gupta, S. K. "Abstractive summarization: An overview of the state of the art." Expert Syst. Appl., vol. 121, pp. 49–65. URL: https://doi.org/10.1016/j.eswa.2018.12.011. doi:10.1016/j.eswa.2018.12.011, **2019.**

[4].Kumarasamy Mani, S. K., Catherine, R., Sinha, V., & Dubey, A. "Ausum: Approach for unsupervised bug report summarization." (p. 11). doi:10.1145/2393596.2393607, **2012.**

[5]. Lotufo, R., Malik, Z., & Czarnecki, K. "Modelling the 'hurried' bug report reading process to summarize bug reports". Empirical Software Engineering Journal, vol. 20, pp. 516– 548. doi:10.1007/s10664-014-9311-2, **2012.**

[6] Rastkar, S., Murphy, G. C., & Murray, G. "Summarizing software artifacts:a case study of bug reports." In Proceedings of the 26th Conference on Program Comprehension ICSE **2010.**

[7] Rastkar, S., Murphy, G. C., & Murray, G. "Automatic summarization of bug reports." IEEE Transactions on Software Engineering, **vol. 40, pp. 366–380, 2014.**

[8] YANG, C.-Z., Cheng-Min, & CHUNG, Y.-H.. "Towards an improvement of bug report summarization using two-layer semantic information." IEICE TRANS. INF. and SYST., **vol. 101, pp. 1743– 1750, 2018.**

[9]. Limsettho, Nachai & Hata, Hideaki & Monden, Akito & Matsumoto, Kenichi. "Automatic Unsupervised Bug Report Categorization," **2014.**

[10]. Beibei Huai, Wenbo Li, Qiansheng Wu, Meiling Wang ":Mining Intentions to Improve Bug Report Summarization." SEKE 2018: **pp. 320-319, 2018.**

AUTHORS PROFILE

Mrs Som Gupta is an MTech in Information Technology from International Institute of Information Technology Bangalore. She has done BE in Computer Science Engineering from Gujarat University. Her schooling is from CBSE Board. She has published 9 research papers with 2 SCI, 2 Scopus and rest all of good International journals.

Dr. S. K. Gupta is presently working as Associate Professor in Comp. Sc. & Engg. Deptt. of BIET, Jhansi, U.P., INDIA. Initially graduating in Computer Science & Engineering from H.B.T.I., Kanpur, U.P., INDIA and then M.E. from M.N.R.E.C., Allahabad, U.P., INDIA and after that completed Ph.D. in Computer Science & Engineering from Bundelkhand University, Jhansi, U.P., INDIA. His area of interest includes image processing and data mining. He has published more than 50 papers.

