Machine Learning Best Practices Discussed on Stack Exchange

German David Martínez Solano
Systems and Computing Engineering Department
Universidad de los Andes
Bogotá, Colombia
gd.martinez@uniandes.edu.co

Mónica Andrea Bayona Latorre
Systems and Computing Engineering Department
Universidad de los Andes
Bogotá, Colombia
ma.bayona@uniandes.edu.co

Mario Linares Vásquez
Systems and Computing Engineering Department
Universidad de los Andes
Bogotá, Colombia
Adviser
m.linaresv@uniandes.edu.co

Anamaria Irmgard Mojica Hanke
Systems and Computing Engineering Department
Universidad de los Andes
Bogotá, Colombia
Co-advisor
ai.mojica10@uniandes.edu.co

***Abstract*—Throughout this paper, we seek to carry out an analysis about which best/good practices of Machine Learning (ML) are discussed in Communities Question Answerings (CQ&As) websites, and establish if these practices are being used in Software Engineering (SE). To achieve this we will follow a series of steps to extract information from different CQ&As communities from Stack Exchange (STE) website, download the users dumps from the selected pages, preprocess it and obtain the relevant information. In addition, we will analyze and collect information about posts that were already tagged, with possible best practices, and a plausible taxonomy of ML best practice. This analyzed will be executed by (i) analyzing which of the best practices in the taxonomy are being used in SE conference studies; (ii) surveying the SE articles authors, in order to understand which good practice have they followed. Subsequently, for the opposite phase, best/good software practices that are used in ML, the goal is review related work and its state of the art.**

***Index Terms*—machine learning, software engineering**

I. Introduction

Currently, there are numerous Communities Question Answering (CQ&A) websites which allow the exchange of knowledge on different topics of interest, in our case, we identify a massive exchange of software knowledge and, with the rise of Machine Learning (ML), many doubts, errors and discussions in these forums have arisen. However, the topic of best/good ML practices has not been explored in depth in these communities and therefore, it is unknown if the recommendations or solutions given in these forums follow good practices.

This paper is made in order to know the ML good practices in the Software Engineering (SE) community, to know if the community is applying them and if Stack Exchange (STE) is making good recommendations in this field. With this study, it is expected that anyone will have at their disposal what are those recommended good practices that should be applied when conducting a study with ML, both for experts and non-experts on the subject.

II. Related work and state-of-the-art

STE is one of the most used CQ&A websites, within its communities is Stack Overflow (STO), the largest and the one most trusted by developers to share and learn [1]. Due to all the knowledge that is exchanged in the forums, these have been a research focus on the SE field. However, from the prior work we analyzed, it is worth noting that none were related to good ML practices in CQ&As. The main topics consisted of trends, impact of STO in a specific field [2]–[5] in the CQ&As, difficulties or challenges presented on specific topics [6]–[8], expertise of the users in STO [6], [9], [10] or discussion about libraries, APIs or frameworks [7], [11]–[13]. Given the nature of our research, we will aim previous studies focused on ML.

Alshangiti et al. [6] analyzed the ML posts and their respective users on STO, and they were able to extract interesting insights from the developers challenges in this field. Some of the most outstanding findings were the lack of ML experts, which was obtained by comparing the ExpertiseRank between ML and Web Development in STO. Moreover, it was also concluded that questions concerning ML take considerably longer to receive an answer than other typical questions; and to summarize, they found using natural language processing that data and feature preprocessing are the most challenging topics in the ML questions.

The study by Islam et al. focused on questions about popular ML libraries (i.e., Caffe [14], H2O [15], Mahout, Keras [16], MLlib, scikit-learn [17], Tensorflow [18], Theano [19], Torch [20] and Weka [21]). They posed four research questions, concerning ML libraries, along the ML pipeline of [22] and answered them using 3,283 STO questions that were manually tagged. Regarding the ML pipeline stages, it was determined that when it comes to libraries that support ML clusters, the

model creation stage presented greater difficulties, followed by data preparation. They also found that type mismatch problems in the input data were present in most libraries, while shape mismatches (errors related to dimensions in the matrix-tensor layers) were more frequent in deep learning libraries. Another finding was that the model creation problems were consistent throughout the study period, while the data preparation problems decreased and increased again at 4 years. In this way they determined that some issues are related to specific periods.

Finally, Han et al. [11] focused on discussions of three Deep Learning (DL) frameworks (i.e., Tensorflow, Theano, Pytorch), on STO and GitHub. They downloaded 26,887 STO posts and 36,330 GitHub pull requests and issues, which they used to identify, through Latent Dirichlet Allocation (LDA), categories of the DL workflow stages. Among the most outstanding findings, is that in both platforms and the three frameworks, the most discussed stages are Model Training and preliminary preparation, and the most popular topic is Error.

In our research, we use STO questions and also posts from other Communities Question Answerings (CQAs) in the STE data dump [23].

III. General and specific objectives

A. *General objective*

Extract and analyze which ML best/good practices are discussed in CQ&As websites and establish if these practices are being used in SE.

B. *Specific objectives*

1) Extract from conference articles what ML good practices are being applied in SE.
2) Create a taxonomy visualization based on the good practices found.
3) Analyze ML good practices in conference papers.
4) Extract and analyze information concerning ML good practices used by papers' authors.
5) Perform validation with experts in the area on good practices of ML they have used in SE.
6) Investigate the state of the art of good software practices in ML.

IV. Execution plan

The work methodology to be followed is the Kanban strategy, which consists of the elaboration of a table or diagram in which three columns of tasks are reflected; pending, in process or completed. We added a fourth column with the tasks that are going to be executed in future weeks, but are not relevant at the moment of creation.

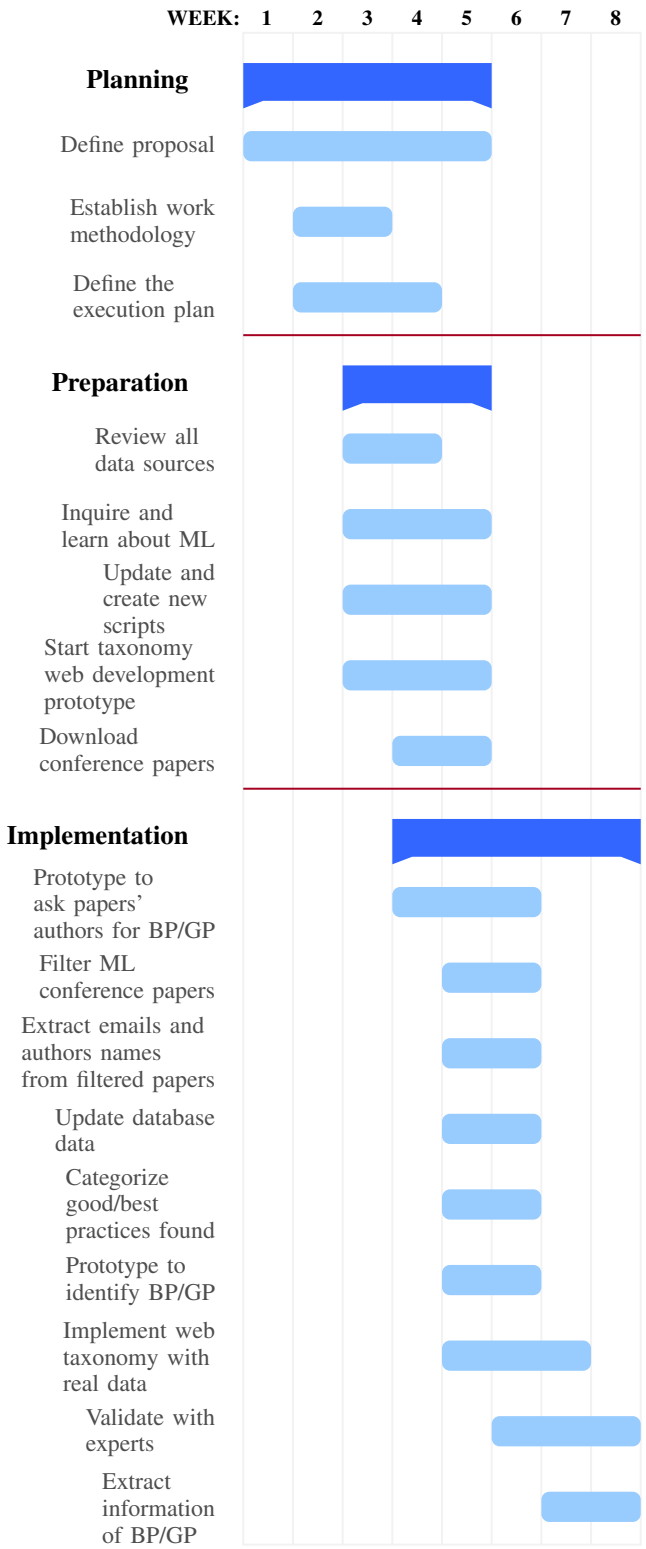
A weekly meeting will be held with the co-advisor to plan new tasks and review the progress in the Kanban board, thus allowing greater control and review of tasks in a faster and easier way, also allowing the collaboration of all members in the tasks.

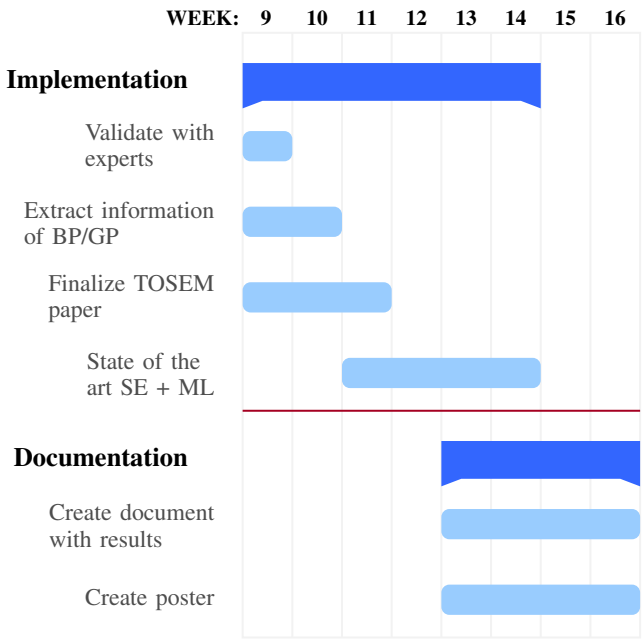
Table I shows the phases and their activities that will be completed throughout the semester to achieve all the project objectives.

Phase	Activities
Planning	Define proposal
	Establish work methodology
	Define the execution plan
Preparation	Review all data sources
	Inquire and learn about ML
	Update and create new scripts
	Start taxonomy web development prototype
	Download conference papers
Implementation	Prototype to ask papers' authors for BP/GP
	Filter ML conference papers
	Extract emails and authors names from filtered papers
	Update database data
	Categorize good/best practices found
	Prototype to identify BP/GP
	Implement web taxonomy with real data
	Validate with experts
	Extract information of BP/GP
	Finalize TOSEM paper
	State of the art SE + ML
Documentation and closure	Create document with results
	Create poster

TABLE I
Planning of activities

The following figure corresponds to a Gantt chart whose objective is to expose the time of dedication planned for different tasks or activities throughout the semester.





V. Expected results

At the end of the project, the taxonomy of good ML practices used in software engineering is expected to be validated with experts in the area and we could determine if these practices are being followed by people in the field. Also, we would know which good practices are being applied in ML conference papers and which other practices are being used by these paper's authors. In addition, it is expected to know the state of the art and documentation regarding good software practices that are applied in ML.

References

[1] StackExchange. 2021. Stack exchange About. https://stackexchange.com/about

[2] A. A. Bangash, H. Sahar, S. Chowdhury, A. W. Wong, A. Hindle, and K. Ali. 2019.What do Developers Know About Machine Learning: A Study of ML Discussionson StackOverflow. In2019 IEEE/ACM 16th International Conference on MiningSoftware Repositories (MSR). 260–264. https://doi.org/10.1109/MSR.2019.00052

[3] Anton Barua, Stephen W. Thomas, and Ahmed E. Hassan. 2014. What aredevelopers talking about? An analysis of topics and trends in Stack Overflow.Empirical Software Engineering19, 3 (June 2014), 619–654. https://doi.org/10.1007/s10664-012-9231-y

[4] Sarah Meldrum, Sherlock A. Licorish, and Bastin Tony Roy Savarimuthu. 2017. Crowdsourced Knowledge on Stack Overflow: A Systematic Mapping Study. InProceedings of the 21st International Conference on Evaluation and Assessment inSoftware Engineering(Karlskrona, Sweden)(EASE'17). Association for Comput-ing Machinery, New York, NY, USA, 180–185. https://doi.org/10.1145/3084226.3084267

[5] Christoph Treude, Ohad Barzilay, and Margaret-Anne Storey. 2011. How Do Pro-grammers Ask and Answer Questions on the Web? (NIER Track). InProceedingsof the 33rd International Conference on Software Engineering(Waikiki, Honolulu,HI, USA)(ICSE '11). Asso-ciation for Computing Machinery, New York, NY, USA,804–807. https://doi.org/10.1145/1985793.1985907

[6] M. Alshangiti, H. Sapkota, P. K. Murukannaiah, X. Liu, and Q. Yu. 2019. Whyis Developing Machine Learning Applications Challenging? A Study on StackOverflow Posts. In2019 ACM/IEEE International Sym-posium on Empirical SoftwareEngineering and Measurement (ESEM). 1–11. https://doi.org/10.1109/ESEM.2019.8870187

[7] M. J. Islam, H. Nguyen, Rangeet Pan, and H. Rajan. 2019. What Do DevelopersAsk About ML Libraries? A Large-scale Study Using Stack Overflow.ArXivabs/1906.11940 (2019)

[8] Pavneet Singh Kochhar. 2016. Mining Testing Questions on Stack Overflow. InProceedings of the 5th International Workshop on Software Mining(Singapore,Singapore)(SoftwareMining 2016). Association for Computing Machinery, NewYork, NY, USA, 32–38. https://doi.org/10.1145/2975961.2975966

[9] S. L. Vadlamani and O. Baysal. 2020. Studying Software Developer Expertiseand Contributions in Stack Overflow and GitHub. In2020 IEEE InternationalConference on Software Maintenance and Evolution (ICSME). 312–323. https://doi.org/10.1109/ICSME46990.2020.00038

[10] B. Yang and S. Manandhar. 2014. Exploring user expertise and descriptive abilityin community question answering. In2014 IEEE/ACM International Conferenceon Advances in Social Networks Analysis and Mining (ASONAM 2014). 320–327.https://doi.org/10.1109/ASONAM.2014.6921604

[11] Junxiao Han, Emad Shihab, Zhiyuan Wan, Shuiguang Deng, and Xin Xia. 2020.What do Programmers Discuss about Deep Learning Frameworks.EmpiricalSoftware Engineering25, 4 (Jul 2020), 2694–2747. https://doi.org/10.1007/s10664-020-09819-6

[12] Y. Hashemi, M. Nayebi, and G. Antoniol. 2020. Documentation of MachineLearning Software. In2020 IEEE 27th International Conference on Software Anal-ysis, Evolution and Reengineering (SANER). 666–667. https://doi.org/10.1109/SANER48275.2020.9054844

[13] Md Ahasanuzzaman, Muhammad Asaduzzaman, Chanchal K. Roy, and Kevin A.Schneider. 2020. CAPS: a supervised technique for classifying Stack Overflowposts concerning API issues.Empirical Software Engineering25, 2 (Mar 2020),1493–1532. https://doi.org/10.1007/s10664-019-09743-4

[14] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long,et al.2014. Caffe: Convolutional Architecture for Fast Feature Embedding.arXivpreprint arXiv:1408.5093(2014)

[15] Erin LeDell and Sebastien Poirier. 2020. H2O AutoML: Scalable AutomaticMachine Learning.7th ICML Workshop on Automated Machine Learning (AutoML)(July 2020). https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_61.pdf

[16] François Chollet and others. 2015. Keras. https://keras.io.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, et al.2011. Scikit-learn: Machine Learning in Python.Journal of Machine Learning Research(2011)

[18] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, et al.2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.https://www.tensorflow.org/ Software available from tensorflow.org.

[19] Rami Al-Rfou, Guillaume Alain, Amjad Almahairi, Christof Angermueller,Dzmitry Bahdanau, et al.2016. Theano: A Python framework for fast com-putation of mathematical expressions.arXiv e-printsabs/1605.02688 (May 2016).http://arxiv.org/abs/1605.02688

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, et al.2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. InAdvances in Neural Information Processing Systems 32, H. Wallach, H. Larochelle,A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates,Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[21] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann,and Ian H. Witten. 2009. The WEKA data mining software: an update.SIGKDDExplorations11, 1 (2009), 10–18

[22] Yufeng Guo. 2017. The 7 Steps of Machine Learning. https://towardsdatascience.com/the-7-steps-of-machine-learning-2877d7e5548e

[23] Stack Exchange Community. 2021. Stack Exchange Data Dump: Stack Exchange,Inc.: Free Download, Borrow, and Streaming.https://archive.org/details/stackexchange