Beyond Accuracy: ROI-driven Data Analytics of Empirical Data

Gouri Deshpande
gouri.deshpande@ucalgary.ca
University of Calgary
Calgary, Canada

Guenther Ruhe
ruhe@ucalgary.ca
University of Calgary
Calgary, Canada

ABSTRACT

Background: The unprecedented access to data has rendered a remarkable opportunity to analyze, understand, and optimize the investigation approaches in almost all the areas of (Empirical) Software Engineering. However, data analytics is time and effort consuming, thus, expensive, and not automatically valuable.

Objective: This vision paper demonstrates that it is crucial to consider Return-on-Investment (ROI) when performing Data Analytics. Decisions on "How much analytics is needed"? are hard to answer. ROI could guide for decision support on the What?, How?, and How Much? analytics for a given problem.

Method: The proposed conceptual framework is validated through two empirical studies that focus on requirements dependencies extraction in the Mozilla Firefox project. The two case studies are (i) Evaluation of fine-tuned BERT against Naive Bayes and Random Forest machine learners for binary dependency classification and (ii) Active Learning against passive Learning (random sampling) for *REQUIRES* dependency extraction. For both the cases, their analysis investment (cost) is estimated, and the achievable benefit from DA is predicted, to determine a break-even point of the investigation.

Results: For the first study, fine-tuned BERT performed superior to the Random Forest, provided that more than 40% of training data is available. For the second, Active Learning achieved higher F1 accuracy within fewer iterations and higher ROI compared to Baseline (Random sampling based RF classifier). In both the studies, estimate on, How much analysis likely would pay off for the invested efforts?, was indicated by the break-even point.

Conclusions: Decisions for the depth and breadth of DA of empirical data should not be made solely based on the accuracy measures. Since ROI-driven Data Analytics provides a simple yet effective direction to discover when to stop further investigation while considering the cost and value of the various types of analysis, it helps to avoid over-analyzing empirical data.

KEYWORDS

Data Analytics, Return-on-Investment, Requirements Engineering, Dependency extraction, BERT, Mozilla

ESEM '20, October 8–9, 2020, Bari, Italy

ACM ISBN 978-1-4503-7580-1/20/10…$15.00
https://doi.org/10.1145/3382494.3422159

ACM Reference Format:
Gouri Deshpande and Guenther Ruhe. 2020. Beyond Accuracy: ROI-driven Data Analytics of Empirical Data. In *ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM) (ESEM '20), October 8–9, 2020, Bari, Italy*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3382494.3422159

1 INTRODUCTION

Return-on-Investment (ROI) is of great interest in engineering and business for arriving at decisions. This is true in Software Engineering (SE) as well. For example, Silverio et al. [16] evaluated cost-benefit analysis for the adoption of software reference architectures for optimizing architectural decision-making. Cleland et al. [8] studied the ROI of heterogeneous solutions for the improvement of the ROI of requirements traceability. Recent data explosion in the form of big data and advances in Machine Learning (ML) have posed questions on the efficiency and effectiveness of these processes that have become more relevant. In this paper, we present a retrospective evaluation of two empirical studies taken from the field of requirements dependency analysis for the benefit of ROI.

Data Analytics in SE (also called "Software Analytics" by Bird et al. [6]) is a term widely used, sometimes with a slightly different meaning. We subsume all efforts devoted to collecting, cleaning, preparing, classifying, analyzing data, and interpreting the results as *Data Analytics (DA)*. In SE, the goal of DA is to provide better insights into some aspects of the software development life-cycle, which could facilitate some form of understanding, monitoring, or improvement of processes, products or projects.

SE is uncertain in various ways. SE is highly human-centric, and processes are not strictly repeatable. The goals and constraints of software development are dynamically changing. Experimentation and DA are inherently arduous under such circumstances. The famous Aristotle [2] is widely attributed with a saying, "It is the mark of an educated mind to rest satisfied with the degree of precision which the nature of the subject admits and not to seek exactness where only an approximation is possible". Figure 1 shows a typical

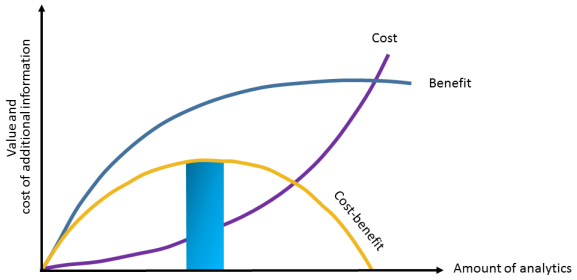


Figure 1: Break-even point from cost-benefit analysis of technology investment.

ROI (cost-benefit) curve of technology usage. Following some phase of increase, the curve reaches saturation, so, beyond that point, further investment does not pay off. We contemplate that a similar behaviour holds true for applying DA. **Our research hypothesis** is that ROI-driven DA helps to determine the break-even point of investment and thus optimizes resources spent in this process.

Paper structure: Section 2 discusses related work. The problem formulation is detailed in Section 3. Section 4 explains the empirical ROI investigation approach for the two problems. A discussion of the applicability of the results is elaborated in Section 5. Finally, Section 6, provides an outlook on future research.

2 RELATED WORK

2.1 ROI Analysis in Software Engineering

Evaluating the profitability of expenditure helps to measure success over a period of time thus takes the guesswork away from the concrete decision-making process. For instance, Erdogmus et al. [15] analyzed the ROI of quality investment to bring its importance in perspective and posed important questions, "We generally want to increase a software products quality because fixing existing software takes valuable time away from developing new software. But how much investment in software quality is desirable? When should we invest, and where?".

Begel & Zimmermann [3] composed a set of 145 questions - based on a survey with more than 200 developers and testers - that are considered relevant for DA at Microsoft. One of the questions:"How important is it to have a software DA team answer this question?", expected answer on a five-point scale (*Essential* to *I don't understand*). Although it provides a sneak peek of the development and testing environments of Microsoft, it does not prove any emphasis on any form of ROI. Essentially, we speculate that the ROI aspect was softened into asking for the perceived subjective importance through this question.

Boehm et al. [7] presented quantitative results on the ROI of Systems Engineering based on the analysis of the 161 software projects in the COCOMO II database. Van Solingen [29] analyzed the ROI of software process improvement and took a macro perspective to evaluate corporate programs targeting the improvement of organizational maturity. Ferrari et al. [17] studied the ROI for text mining and showed that it has not only a tangible impact in terms of ROI but also an intangible benefits - which occur from the investment in the knowledge management solution that is not directly translated into returns, but that must be considered in the process of judgment to integrate the financial perspective of analysis with the non-financial ones. A lot of benefits occurring from the investment in this knowledge management solution are not directly translated into returns, but they must be considered in the process of judgment to integrate the financial perspective of analysis with the non-financial ones.

Ruhe and Nayebi [23] proposed the *Analytics Design Sheet* as a means to sketch the skeleton of the main components of the DA process. The four-quadrant template provides direction to brainstorm candidate DA methods and techniques in response to the problem statement and the data available. In its nature, the sheet is qualitative. ROI analysis goes further and adds a quantitative perspective for outlining DA.

2.2 Empirical Analysis for Requirements Dependency Extraction

The extraction of dependencies among requirements is an active field of SE research. The practical importance of the topic was confirmed by our survey [13]. More than 80% of the participants agreed or strongly agreed that (i) dependency type extraction is difficult in practice, (ii) dependency information has implications on maintenance, and (iii) ignoring dependencies has a significant ill impact on project success.

In the recent past, many empirical studies have explored diverse computational methods that used natural language processing (NLP) [10] [24], semi-supervised technique [12], hybrid techniques [11] and deep learning [18]. However, none of the approaches considered ROI to decide among techniques and the depth and breadth of their execution level.

3 CONCEPTUAL FRAMEWORK FOR ROI-DRIVEN DATA ANALYTICS

Different models exist that provide guidance to perform DA. Wieringa [30] provides a checklist for what he calls the design cycle and the empirical cycle. In this study, we use the term *Scoping* for defining the problem and the analysis objectives. Scoping also means defining the boundaries that help to exclude non-essential parts of the investigation. Analysis of the projected *Return-on-Investment (ROI)* serves as an input for scoping.

3.1 Research Question

DA follows a resource and computation-intensive process constituting data gathering and processing components that are the non-trivial proportion of the total research cost. Thus, it is essential to account for these to compute the overall cost-benefit and optimize it further.

Our aim is to look at DA for empirical studies retrospectively (already conducted studies in the past). In particular, we are interested in Requirements Dependency Analysis (RDA) based studies. Through this research, we define and validate the principal concepts needed for ROI-driven DA. Our research question is:

RQ: What are the benefits of ROI-driven Data Analytics in the studies focusing on Requirements Dependency Analysis?

Justification: As for any investment, it is most important to know how much is enough. There is no incentive to invest in analytics just for the sake of performing some analysis. Although one cannot claim exactness from this, it is worthwhile to get some form of guidance on where (which techniques) and how far (how much of it) one should go. To make the analysis concrete, we have selected RDA as the area of our specific investigations.

3.2 Cost Factors

Data processing is an umbrella term used to combine data collection (36), pre-processing (??) and labeling (;) under one hood, each one of which is a cost component. However, not all costs are fixed and some vary based on the solution approach used to tackle any decision problem. For example, supervised Machine Learning (ML) requires a large amount of annotated data, to begin with,

Table 1: Parameters used for ROI computation

	Symbol	Meaning	Unit
Cost	36	Data gathering time	Minutes
	??	Pre-processing time	Minutes
	4	Evaluation time	Minutes
	;	Labeling time	Minutes
	A4B>DA24	Human resource cost	$ per hour
Bene t	A4F0A3	Value per TP	$
	?4=0;C~	Penalty per FN	$
	1 8C4A0C8>=	F1 di erence	Number
	%+0;D4	Projected value per 1% F1 improvement	$
Others		#Human resources	Number
	# CA08=	Size of the training set	Number
	# C4BC	Size of the test set	Number
	#	# CA08= + # C4BC	Number

whereas Active Learning acquires these annotations over a period of time in iterations until a stopping condition for classi cation operation is reached [25]. Additionally, there is a cost associated with modeling and evaluation (4).

3.3 Value Factors

The value returns or bene ts are de ned based on the needs of the decision problem. In the context of dependency extraction, the bene t could be modeled in terms of the ability of the ML model to identify a larger number of dependencies correctly (higher # of True Positives TP: A4F0A3) while limiting misclassi cation (reduced # of False Negatives FN: ?4=0;C). Conversely, the bene t could also be determined based on the net value %+(0;D4 of change of accuracy (1 8C4A0C8>=) in every iteration, especially when using Active Learning. Table 1 lists the relevant cost components and their corresponding units. These will be utilized to compute the '$ later for the two di erent problems in Section 4.4.

3.4 ROI

To determine the ROI, we follow the simplest form of its calculation relating to the di erence between 4=4582 and >BC to the amount of >BC. Both 4=4582 and >BC are measured as human e ort in person hours.

$$'\$ = {}^{1}4{=}4582C >BC^{\circ} \bullet >BC \quad (1)$$

Costa et al. [9] distinguished the hard ROI from the soft ROI. The former refers to the direct additional revenue generated and cost savings. The latter improved productivity, customer satisfaction, technological leadership, and e ciencies.

4 ROI OF TECHNIQUES FOR REQUIREMENTS DEPENDENCY ANALYSIS

We have selected the area of requirements dependency analysis (RDA) to illustrate and initially validate our former conceptual framework. In what follows, we introduce the key terms needed to formulate two Empirical Analysis Studies called EAS 1 resp. EAS 2.

4.1 Problem statement

Following are the de nitions of dependency types that are used to state the two studies. For a set of requirements and a pair of requirements 1A•Bn ' '

1) An INDEPENDENT relationship is de ned as the absence of any form of relationship between a pair of requirements.
2) A DEPENDENT relationship is de ned as the complement set of INDEPENDENT. i.e., there exists at least one type of the dependency types such as REQUIRES, SIMILAR, OR, AND, XOR, value synergy, e ort synergy etc. between A and B.
3) REQUIRES is a special form of DEPENDENT relationship. If A requires B, or B requires A, then, A and B are in a REQUIRES relationship
4) OTHER type of dependency is when 1A•Bis DEPENDENT and the dependency type is not REQUIRES (could be any of the other dependency types mentioned in (2))

Problem 1- Binary requirements dependency extraction: For a given set of requirements and their textual description, the binary requirements dependency extraction problem aims to classify each pair (r,s)n ' ' as DEPENDENT or INDEPENDENT

Problem 2- Speci c requirements dependency extraction of the type REQUIRES: For a given set of requirements and their textual description, the REQUIRES dependency extraction problem aims to classify for each pair (r,s)' ' if they are in a' &* ' (relationship.

4.2 Empirical Analysis Studies (EAS)

In this section, we formulate two Empirical Analysis Studies, EAS 1 and EAS 2, to investigate the two problems explained above. We aim to analyze and compare Bidirectional Encoder Representations from Transformers (BERT), and Active Learning (AL), both proven to be of interest in general and pre-evaluated for their applicability to the stated problems, with traditional ML. For the two studies, we examine the (F1) accuracy and the ROI of the whole process of DA.

EAS 1: We compare two supervised classi cation algorithms: Naive Bayes (NB) and Random Forest (RF) - ML algorithms successfully and prominently used for text classi cation [19] in the past, with a ne-tuned BERT model [14]. The analysis was performed for an incrementally growing training set size to capture its impact on F1 accuracy and ROI.

BERT (Bidirectional Encoder Representations from Transformers) [14] is a recent technique published by researchers from Google. BERT is applying bidirectional training of Transformer, a popular attention model, to language modeling, which claims to be state-of-the-art for NLP tasks. In this study scenario, we explore the question, How does ne-tune BERT compare with traditional algorithms on an economical scale? by comparing models' e ectiveness with incurred ROI.

EAS 2: Random sampling (Passive Learning) randomly selects a training set - referred to as Baseline in the rest of the paper. Active Learning selects the most informative instances using various sampling techniques such as MinMargin and LeastCon dence [25]. We compare Baseline with AL using RF as a classi er for this scenario. The analysis was done by adding a few training samples in every iteration concurrently to classify the unlabeled instances.

Active Learning(AL) is a ML method that guides a selection of the instances to be labeled by an oracle (e.g., human domain expert or a program) [25]. While this mechanism has been proven to positively address the question, Can machines learn with fewer labeled training instances if they are allowed to ask questions? , through this exploration, we try to answer the question, Can machines learn more economically if they are allowed to ask questions? [26].

4.3 Data

The online bug tracking system Bugzilla [20] is widely used in open-source software development. New requirements are logged into these systems in the form issue reports [7, 4] which help software developers to track them for e ective implementation [9], testing, and release planning. In Bugzilla, feature requests are a speci c type of issue that is typically tagged as enhancement [11]. We retrieved these feature requests or requirements from Firefox and exported all related elds such as Title, Type, Priority, Product, Depends_on, and See_also.

Data collection: Collecting data from Bugzilla was a substantial e ort that was carried out in multiple rounds. We collected 3,704 enhancements from Firefoxusing REST API through a python script such that each one of the enhancements considered for retrieval is dependent on at least another one in the dataset. The data spanned from 08/05/2001 to 09/08/2019.

Data preparation: The complete data was analyzed to eliminate special characters and numbers. Then dependent requirement pairs were created based on the depends_on (interpreted as REQUIRES dependency) eld information for each one of the enhancements. Requirements with no dependency between them were paired to generate INDEPENDENT class dataset. Further, sentence pairs that had fewer than three words in them were ltered out resulting in 3,373 REQUIRES, 219 OTHER and 21,358 INDEPENDENT pairs.

Pre-processing and feature extraction: The data was rst processed to eliminate stop words and then lemmatized following the traditional NLP pipeline [1]. For supervised and AL ML, we used the Bag Of Words (BOW) [22] feature extraction method, which groups textual elements as tokens. For applying BERT, we retained sentence pairs in their original form (without stop word removal and lemmatization).

Classi ers: For both NB and RF, the data was split into train and test (80:20) and balanced between classes. Also, hyper-parameter tuning was performed and the results for 10-fold cross-validation were computed, followed by testing (on unseen data).

To ne-tune the BERT model, we used NextSentencePrediction [1] on a sentence pair classi cation pre-trained BERT model, and further ne-tuned it for the RDA speci c dataset on Tesla K80 GPU on Google Colab [2].

4.4 ROI Modeling

4.4.1 EAS 1. The classi cation algorithms such as RF and NB, have been explored in NLP based SE problems. These algorithms are driven by the feature extraction aspect to a great extent. Thus, could in uence their e ectiveness on classi cation outcomes. However, feature extraction is problem speci c and incurs substantial cost and access to domain expertise.

On the other hand, BERT eliminates the need for feature extraction since it is a language model based on deep learning. BERT, pre-trained on a large text corpus, can be ne-tuned on speci c tasks by providing only a small amount of domain-speci c data.

In this empirical analysis, we conducted classi cation by utilizing a fraction of the whole dataset for training and testing for a small xed data set. This was repeated by slowly increasing the fraction of the training set and results were captured.

During every classi cation, >BC and 4=4580 were computed using various parameters explained in Table 1. >BC is the sum of the data processing costs (36, ??, 4, ;°•60) (in hours) for a fraction (N%) of training set. This is further translated into dollar cost based on hourly charges (A4B>DA24) of human resources.

$$>BC = \#\% \frac{^{1}\,36,\ ??,\ 4,\ ;^{o}}{60} \quad A4B>DA24 \tag{2}$$

'4CDA computations for RDA, assumes reward (A4F0A3) for identifying the dependent requirements (TP) while penalizing (?4=0;C~) instances that were falsely identi ed as independent (FN).

$$4=4580 \%\ A4F0A3\ \#\ ?4=0;C\sim \tag{3}$$

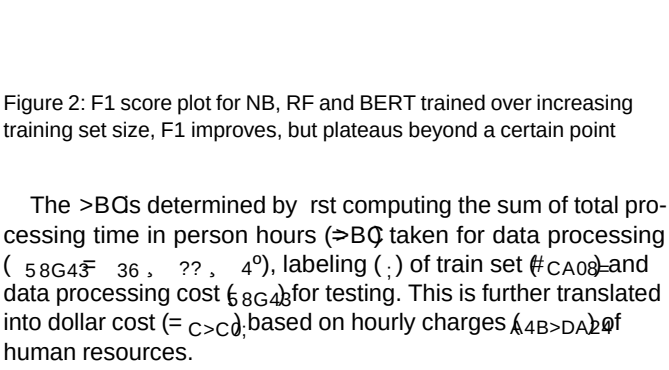
Table 2: Parameter settings for the two empirical analysis scenarios

Parameters	Values
58G43= 36, ??, 4	1 min/sample
;	0.5 min/sample
A4B>DA24	$400/hr
	1
#	4,586
A4F0A3	$500/TP
?4=0;C~	$500/FN
1 8C4A0C8>= %+0;D4	= 2DA ?A4E $10,000 per percent F1 improvement

4.4.2 EAS 2. In this empirical analysis, we compared AL with a traditional random sampling based classi cation Baseline using the RF ML algorithm.

Beginning with 60 training samples of each class (REQUIRES, INDEPENDENT and OTHER), we developed multi-class classi ers for both AL and Baseline for this empirical study scenario. When AL used MinMargin sampling technique [3] to identify 20 [4] most uncertain instance (requirement pair) for oracle to label, baseline randomly selected 20 instances and added to the training set along with their label, thus, kept the two approaches comparable in all the 20 iterations. Since data is already labeled, for AL, we pretend they are unlabeled until queried and labeled by a simulated oracle in this scenario.

[1] https://huggingface.co/transformers/model_doc/bert.html#bertfornextsentenceprediction
[2] https://colab.research.google.com/
[3] MinMargin sampling technique performed well compared to Least Con dence and Entropy thus, we utilized MinMargin for this study
[4] The tests were performed with#samples = 10, 15 and 20. In this study, we will discuss results related to #samples=20



(a) F1 vs ROI for Random Forest (b) F1 vs ROI for Fine tuned BERT

Figure 3: Empirical Analysis Scenario 1 (EAS 1)

Figure 2 provides the accuracy only view and shows that F1 gradually increases with the increasing training size for the three ML algorithms: NB, RF, and BERT. However, all three ML algorithms reach a saturation towards larger training set sizes. While BERT performed exceptionally well when training set size exceeded 42%, it could have been ideal to pre-determine How much training is enough? . Thus we selected the top two classi ers (Figure 2): BERT and RF and applied the monetary values (Table 2) for the various cost and bene t factors de ned in Table 1 and computed the ROI.

Figure 3a and 3b show the results for RF and BERT, respectively. The ROI behaviour is not monotonous and peaks for both cases. Although RF classi cation achieved the highest ROI with just 20% of training set and accuracy of F1 = 0.7, highest F1 value of 0.75 was achieved along with the lowest ROI of 4.7.

For RF classi cation and applying ROI arguments, learning can be stopped with 20% of the training set.

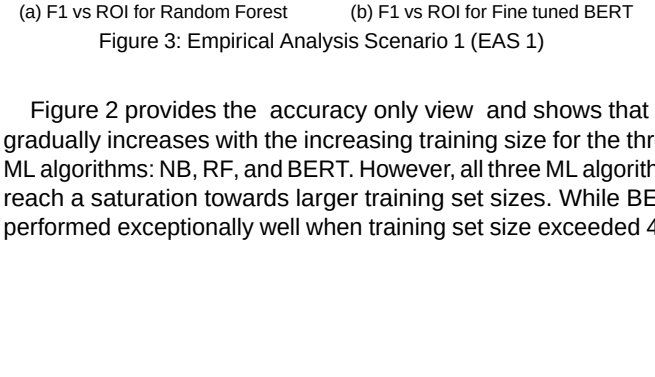
Now looking at BERT classi cation, the best ROI-driven results: F1 = 0.84 and an ROI = 8.43, were achieved with the 60% training set. Although F1 rose to 0.9 with 70% training set size, ROI dropped to 7.27. For the recommendation of 20% of training set size, ROI has a local optimum. BERT in general performs well on the F1, however, is it worth the ROI? needs to be explored.

For training set sizes of at least 40% of the size of the whole set, BERT performed better than RF in terms of both accuracy and ROI.

5.2 EAS 2

We analyzed the ROI for Baseline against AL for classifying the REQUIREDClass. The results are shown in Figure 4a and Figure 4b. Similar to EAS 1, we applied the values from Table 2 and equations (4) and (5) to compute cost and bene t at every iteration for both the approaches. For the Baseline approach, ROI peaked at 3.2 and F1 = 0.6, in the very 2nd iteration. Onwards, ROI drastically decreased which indicated lesser value for increasing training set by random sampling (Baseline) method.

Similar behavior was observed for the AL approach. shown in Figure 4b. The peak here was after three iterations with values ROI = 4.5 and F1 = 0.8.



(a) F1 vs ROI for Baseline (b) F1 vs ROI for AL

Figure 4: Empirical Analysis Scenario 2 (EAS2)

Both Baseline and AL showed the best ROI performance in the early iterations. Higher F1 accuracy needs additional human resources and reduces the ROI.

6 DISCUSSION

For the problem of RDA, we explored the potential value of ROI-driven decisions. When chasing higher accuracy, there is a risk of