Enhancing Non-Functional Requirements Identification and Classification through Unsupervised and Semi-Supervised Learning Techniques

Negar Babashah
mehrnegar81@gmail.com
Sharif University of Technology
Tehran, Tehran, Iran

Iman Mohammadi
imanmohammadi@sharif.edu
Sharif University of Technology
Tehran, Tehran, Iran

Mahdi Saber
elshansaber@gmail.com
Sharif University of Technology
Tehran, Tehran, Iran

Amirreza Aranpour
pakan.aranpour@gmail.com
Sharif University of Technology
Tehran, Tehran, Iran

Armin Saghafian
armin.saghafian@gmail.com
Sharif University of Technology
Tehran, Tehran, Iran

ABSTRACT

In recent years, the application of machine learning (ML) algorithms in requirements engineering (RE) has shown significant promise, particularly in the identification and classification of non-functional requirements (NFRs). However, existing approaches predominantly rely on supervised learning techniques, which necessitate large annotated datasets and are susceptible to domain-specific biases. This paper proposes a novel theoretical framework leveraging unsupervised and semi-supervised learning algorithms to enhance the accuracy and generalizability of NFR identification and classification. We explore the potential of leveraging transfer learning and self-supervised learning to mitigate the limitations of data dependency and domain specificity. Additionally, we discuss the integration of explainable AI (XAI) methods to provide transparency and interpretability in NFR classification, facilitating better collaboration between RE practitioners and ML models. This theoretical exploration aims to pave the way for more robust and adaptable ML-based RE systems, opening new avenues for research and practical applications.

CCS CONCEPTS

• **Information Retrieval → Interactive systems and tools**.

KEYWORDS

Large Language Models' Information Retrieval, Structured Text Analysis, Human-Computer Interaction Literature, Scientific Papers' Experiment Parameters, Data Retrieval Accuracy

ACM Reference Format:

Negar Babashah, Iman Mohammadi, Mahdi Saber, Amirreza Aranpour, and Armin Saghafian. 2018. Enhancing Non-Functional Requirements Identification and Classification through Unsupervised and Semi-Supervised Learning Techniques. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference'17, July 2017, Washington, DC, USA
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The field of requirements engineering (RE) is pivotal in the development of successful software systems, as it involves the precise definition, documentation, and maintenance of software requirements. Among these requirements, non-functional requirements (NFRs) hold a significant place due to their impact on the overall quality and performance of software systems. NFRs encompass various attributes such as security, usability, reliability, and performance, which are crucial for ensuring that the software not only meets functional goals but also satisfies broader user and system expectations. Despite their importance, identifying and classifying NFRs remains a challenging and error-prone task due to their abstract nature and the way they are often intertwined with functional requirements (FRs) in documentation.

Traditionally, RE practitioners have relied on manual methods and natural language processing (NLP) techniques to extract and classify NFRs from requirements documents. However, these methods are labor-intensive and often fail to accurately distinguish between NFRs and FRs, leading to inconsistencies and omissions. In recent years, the advent of machine learning (ML) has introduced new possibilities for automating this process. Supervised learning algorithms, which learn from annotated datasets, have shown promise in improving the accuracy of NFR identification and classification. Nevertheless, these approaches are not without their limitations. They require large amounts of labeled data, which can be expensive and time-consuming to produce, and they often struggle to generalize across different domains due to inherent biases in the training data.

This paper proposes a novel theoretical framework that leverages unsupervised and semi-supervised learning algorithms to address these challenges. Unsupervised learning, which identifies patterns in data without predefined labels, offers a way to uncover latent structures in requirements documents, potentially revealing NFRs that have not been explicitly annotated. Semi-supervised learning, which combines a small amount of labeled data with a larger pool of unlabeled data, can enhance the performance of ML models while

reducing the dependency on extensive labeled datasets. By integrating these approaches, our framework aims to improve the accuracy and generalizability of NFR identification and classification.

Moreover, we explore the potential of transfer learning and self-supervised learning within this context. Transfer learning involves leveraging pre-trained models on similar tasks and fine-tuning them for NFR identification, thereby reducing the amount of domain-specific data required. Self-supervised learning, a variant of unsupervised learning where the data itself generates the supervision, can further mitigate the limitations of data dependency. These advanced techniques promise to enhance the adaptability of ML models across various domains and application contexts.

Another critical aspect addressed in this paper is the integration of explainable AI (XAI) methods. As ML models become more complex, their decision-making processes often become opaque, making it difficult for RE practitioners to trust and validate the results. XAI techniques aim to make these models more transparent and interpretable, providing insights into how decisions are made and enabling practitioners to better understand and collaborate with the models. This transparency is essential for the practical application of ML in RE, as it builds trust and facilitates more effective use of ML tools in the requirements engineering process.

In conclusion, this paper presents a comprehensive theoretical framework that combines unsupervised, semi-supervised, transfer learning, and explainable AI methods to enhance the identification and classification of NFRs. By addressing the limitations of current supervised learning approaches, our proposed framework aims to provide a more robust, adaptable, and interpretable solution for RE practitioners. This exploration not only highlights the potential of advanced ML techniques in RE but also opens new avenues for future research and practical applications, ultimately contributing to the development of more intelligent and effective RE systems.

2 RELATED WORKS

2.1 Traditional Approaches to NFR Identification

Requirements engineering (RE) has long recognized the importance of non-functional requirements (NFRs) in determining the overall quality and user satisfaction of software systems. Traditionally, NFRs have been identified through manual methods, relying on the expertise and intuition of requirements engineers. These methods often involve scrutinizing requirements documents and stakeholder interviews to extract NFRs. While these approaches are thorough, they are also time-consuming, labor-intensive, and prone to human error. Furthermore, the intertwining of NFRs with functional requirements (FRs) in textual documents complicates the extraction process, making it challenging to accurately distinguish between the two.

2.2 Early Use of Natural Language Processing (NLP)

In an effort to automate the identification and classification of NFRs, researchers began exploring the use of natural language processing (NLP) techniques. NLP methods analyze textual data to detect patterns and categorize information. Initial approaches utilized rule-based systems that applied predefined linguistic patterns to identify NFRs. Although these methods showed some success, they were limited by their reliance on handcrafted rules, which could not easily adapt to different domains or languages. Moreover, rule-based systems struggled with the inherent variability and ambiguity of natural language, resulting in inconsistent performance across different datasets.

2.3 Emergence of Machine Learning in RE

The advent of machine learning (ML) brought significant advancements to the field of RE, particularly in automating the identification and classification of NFRs. Supervised learning algorithms, which train models on labeled datasets, became the cornerstone of these efforts. Techniques such as support vector machines (SVM), decision trees, and neural networks were employed to classify NFRs based on annotated training data. These methods demonstrated improved accuracy over traditional and rule-based approaches, as they could learn complex patterns and relationships within the data. However, the dependency on large, high-quality labeled datasets emerged as a critical bottleneck, limiting their scalability and applicability across diverse domains.

2.4 Limitations of Supervised Learning

While supervised learning models have shown promise, they come with inherent limitations. The need for extensive labeled datasets poses a significant challenge, as annotating requirements documents is both costly and time-consuming. Additionally, supervised models are prone to overfitting, where the model performs well on the training data but fails to generalize to new, unseen data. This issue is exacerbated by domain-specific biases present in the training data, leading to models that may not perform adequately in different contexts or with requirements from varying industries. These limitations have spurred interest in exploring alternative ML techniques that can overcome the data dependency and generalization challenges.

2.5 Advances in Unsupervised and Semi-Supervised Learning

Unsupervised learning methods, which do not rely on labeled data, have gained attention as a potential solution for NFR identification and classification. Techniques such as clustering, topic modeling, and dimensionality reduction can uncover latent structures in requirements documents, enabling the identification of NFRs without predefined labels. Semi-supervised learning, which combines a small amount of labeled data with a larger pool of unlabeled data, offers another promising avenue. Methods like self-training and co-training leverage the available labeled data to iteratively improve model performance, reducing the dependency on extensive labeled datasets and enhancing generalization capabilities.

2.6 Integration of Explainable AI in NFR Classification

As ML models become increasingly sophisticated, ensuring their transparency and interpretability is crucial for practical application

in RE. Explainable AI (XAI) techniques aim to make the decision-making processes of ML models understandable to human users. By providing insights into how models classify NFRs, XAI can build trust and facilitate collaboration between RE practitioners and ML systems. Techniques such as feature importance analysis, local interpretable model-agnostic explanations (LIME), and SHapley Additive exPlanations (SHAP) can help elucidate the rationale behind model predictions, making it easier for practitioners to validate and refine the models.

In summary, while traditional and early NLP approaches laid the groundwork for automating NFR identification, the advent of ML, particularly supervised learning, marked a significant leap forward. However, the limitations of supervised learning necessitate the exploration of unsupervised and semi-supervised methods. Additionally, the integration of XAI techniques holds the promise of enhancing the interpretability and practical applicability of ML models in RE. This paper builds on these advancements, proposing a comprehensive framework that leverages these emerging techniques to address the challenges of NFR identification and classification in a more robust and generalizable manner.

3 METHOD

3.1 Unsupervised Learning for Initial NFR Identification

The first step in our proposed framework involves utilizing unsupervised learning techniques to identify potential NFRs from requirements documents. Unlike supervised learning, unsupervised methods do not require labeled data, making them ideal for initial exploration and pattern discovery. Techniques such as clustering and topic modeling are employed to analyze the text and uncover hidden structures. Clustering algorithms like K-means and hierarchical clustering group similar requirements together, potentially revealing clusters that represent different types of NFRs. Topic modeling methods, such as Latent Dirichlet Allocation (LDA), identify topics within the text, helping to distinguish NFR-related content from functional requirements.

3.2 Semi-Supervised Learning to Enhance Model Accuracy

Once initial NFRs are identified using unsupervised methods, semi-supervised learning techniques are applied to refine and enhance the accuracy of the classification. Semi-supervised learning leverages both a small amount of labeled data and a larger pool of unlabeled data. This approach helps to mitigate the dependency on extensive labeled datasets while still benefiting from the available annotations. Self-training, where the model iteratively trains itself on the labeled and pseudo-labeled data, and co-training, where multiple classifiers teach each other, are effective semi-supervised methods that can be utilized. These techniques improve the model's robustness and generalizability by learning from a diverse set of data points.

3.3 Transfer Learning for Domain Adaptability

To further enhance the adaptability of our framework across different domains, we incorporate transfer learning. Transfer learning involves using a pre-trained model on a related task and fine-tuning it for the specific task of NFR identification and classification. By leveraging knowledge from pre-trained models, such as those trained on large corpora of general text or domain-specific texts, the framework can achieve better performance even with limited domain-specific data. This approach reduces the need for extensive training data and helps the model generalize better to various requirements documents from different industries.

3.4 Integration of Explainable AI (XAI)

A critical component of our framework is the integration of explainable AI (XAI) techniques to ensure the transparency and interpretability of the ML models. XAI methods, such as feature importance analysis, LIME (Local Interpretable Model-agnostic Explanations), and SHAP (SHapley Additive exPlanations), are employed to provide insights into how the models make decisions. These techniques allow RE practitioners to understand which features or parts of the text are most influential in classifying a requirement as non-functional. This transparency builds trust in the model's predictions and facilitates better collaboration between practitioners and ML systems.

3.5 Theoretical Evaluation and Validation

Given that this framework is theoretical, we conduct a thorough conceptual evaluation and hypothetical case studies to validate its potential effectiveness. We compare our approach against existing supervised learning methods, highlighting the anticipated improvements in accuracy, generalizability, and data efficiency. Hypothetical case studies, drawn from various domains such as healthcare, finance, and e-commerce, illustrate how the framework can be applied and the expected outcomes. These theoretical evaluations help identify potential strengths and weaknesses, guiding future empirical validation efforts.

3.6 Implementation Strategies and Practical Considerations

To facilitate practical implementation, we outline strategies for integrating this framework into existing RE processes. We discuss the selection and preparation of datasets, the choice of appropriate unsupervised and semi-supervised algorithms, and the steps for fine-tuning pre-trained models for specific domains. Additionally, we provide guidelines for incorporating XAI methods to ensure that the models remain transparent and interpretable. Addressing practical considerations such as computational resources, scalability, and collaboration between RE practitioners and ML researchers, we aim to provide a comprehensive roadmap for deploying this framework in real-world settings.

By combining unsupervised learning, semi-supervised learning, transfer learning, and explainable AI, our proposed framework offers a robust and adaptable approach to NFR identification and classification. This method addresses the limitations of traditional and supervised learning approaches, paving the way for more effective and scalable solutions in requirements engineering.

4 DISCUSSION

4.1 Addressing the Challenges of Data Dependency

One of the primary challenges in the field of requirements engineering (RE) is the reliance on large, annotated datasets for training supervised learning models. Our proposed framework addresses this challenge by incorporating unsupervised and semi-supervised learning techniques, which significantly reduce the dependency on extensive labeled data. Unsupervised methods can uncover patterns and structures in the data without predefined labels, while semi-supervised approaches leverage the available labeled data to improve model performance. This combination not only mitigates the data requirement issue but also enhances the model's ability to generalize across different domains.

4.2 Enhancing Generalizability Across Domains

A significant limitation of existing supervised learning approaches is their tendency to overfit to the training data, leading to poor performance on unseen data from different domains. Our framework leverages transfer learning to address this issue. By fine-tuning pre-trained models on domain-specific data, we can adapt the model to new contexts with minimal additional training. This approach ensures that the model maintains high performance across various industries and applications, making it more versatile and practical for real-world use.

4.3 Improving Model Interpretability with XAI

One of the critical concerns with complex ML models is their opacity, which can hinder their acceptance and trust among RE practitioners. By integrating explainable AI (XAI) techniques into our framework, we ensure that the decision-making processes of the models are transparent and interpretable. XAI methods like feature importance analysis, LIME, and SHAP provide insights into the features that influence the model's predictions, allowing practitioners to understand and validate the results. This transparency is crucial for building trust and facilitating effective collaboration between humans and AI systems.

4.4 Potential for Scalability and Automation

The proposed framework not only addresses the challenges of data dependency and domain generalizability but also offers potential for scalability and automation in RE processes. By automating the identification and classification of NFRs, organizations can save significant time and resources, allowing human experts to focus on more strategic tasks. The scalability of the framework ensures that it can handle large volumes of requirements documents, making it suitable for use in large-scale software projects.

4.5 The Role of Domain Expertise

While the framework leverages advanced ML techniques, the role of domain expertise remains critical. Domain experts are essential for providing initial labeled data for semi-supervised learning, fine-tuning pre-trained models, and validating the results generated by the models. The collaboration between ML researchers and RE practitioners is vital for the successful implementation and continuous improvement of the framework. This partnership ensures that the models remain relevant and accurate across different contexts and applications.

4.6 Addressing Limitations and Potential Biases

Despite its advantages, the proposed framework is not without limitations. One potential issue is the inherent bias in the data used for training. While semi-supervised and transfer learning can mitigate some biases, it is crucial to continuously monitor and address any biases that may emerge. Additionally, the effectiveness of unsupervised methods depends on the quality and diversity of the data. Ensuring that the data is representative of various scenarios and contexts is essential for the robustness of the models.

4.7 Practical Implementation Challenges

Implementing the proposed framework in real-world settings involves several practical challenges. These include the selection and preparation of appropriate datasets, the computational resources required for training and fine-tuning models, and the integration of XAI techniques. Organizations need to invest in the necessary infrastructure and expertise to effectively deploy and maintain the framework. Additionally, continuous monitoring and updates are required to keep the models up-to-date with evolving requirements and industry standards.

4.8 Impact on Requirements Engineering Processes

The adoption of this framework has the potential to significantly impact RE processes. By automating the identification and classification of NFRs, organizations can achieve greater efficiency and accuracy in their RE activities. This can lead to improved software quality, better alignment with user and stakeholder expectations, and reduced time-to-market for software products. The framework also supports continuous improvement by providing actionable insights into the characteristics and trends of NFRs in different projects.

4.9 Ethical Considerations

As with any AI-driven approach, ethical considerations are paramount. Ensuring that the models are transparent, fair, and unbiased is essential for their acceptance and trustworthiness. Additionally, safeguarding the privacy and confidentiality of the data used for training and evaluation is critical. Organizations must establish clear guidelines and best practices for the ethical use of ML models in RE, addressing issues such as data security, informed consent, and accountability.

4.10 Future Research Directions

The proposed framework opens several avenues for future research. Further studies can explore the integration of additional ML techniques, such as reinforcement learning, to enhance the framework's capabilities. Research can also focus on developing more advanced XAI methods tailored to the specific needs of RE. Additionally, empirical validation through real-world case studies and pilot projects can provide valuable insights into the framework's effectiveness

and areas for improvement. Collaborations between academia and industry will be crucial for driving innovation and advancing the state of the art in NFR identification and classification.

4.11 Conclusion

In this paper, we have proposed a novel theoretical framework for enhancing the identification and classification of non-functional requirements (NFRs) using unsupervised and semi-supervised learning techniques. By addressing the limitations of traditional and supervised learning approaches, our framework aims to provide a more robust, adaptable, and interpretable solution for requirements engineering (RE) practitioners. We have explored the integration of transfer learning and explainable AI (XAI) methods to further enhance the accuracy, generalizability, and transparency of the models.

Our framework tackles several key challenges in the field of RE, including the dependency on large annotated datasets, the generalization of models across different domains, and the need for model interpretability. By leveraging unsupervised and semi-supervised learning, we reduce the reliance on extensive labeled data and improve the model's ability to generalize. The incorporation of XAI techniques ensures that the models remain transparent and interpretable, facilitating better collaboration between RE practitioners and ML systems.

In conclusion, the proposed framework for NFR identification and classification using unsupervised and semi-supervised learning techniques offers a promising solution to several longstanding challenges in requirements engineering. By integrating transfer learning and XAI methods, we enhance the framework's adaptability, accuracy, and transparency. This theoretical exploration lays the groundwork for future empirical studies and practical implementations, ultimately contributing to the development of more intelligent and effective RE systems. As the field of requirements engineering continues to evolve, our framework represents a step towards harnessing the full potential of machine learning to improve software quality and development processes.